NISTIR 5237

# Report of the ARPA/NIST Workshop on Performance Evaluation of Unmanned Ground Vehicle Technologies

**Prepared by**

**Martin Herman**

Sponsored by

Advanced Research Projects Agency
Software and Intelligent Systems
Technology Office

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Robot Systems Division
Building 220, Room B124
Gaithersburg, MD 20899



NIST

NISTIR 5237

# Report of the ARPA/NIST Workshop on Performance Evaluation of Unmanned Ground Vehicle Technologies

**Prepared by**

**Martin Herman**

Sponsored by

Advanced Research Projects Agency
Software and Intelligent Systems
Technology Office

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Robot Systems Division
Building 220, Room B124
Gaithersburg, MD 20899

August 1993


**U.S. DEPARTMENT OF COMMERCE**
**Ronald H. Brown, Secretary**

**TECHNOLOGY ADMINISTRATION**
**Mary L. Good, Under Secretary for Technology**

**NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY**
**Arati Prabhakar, Director**

# Report of the ARPA/NIST Workshop on Performance Evaluation of Unmanned Ground Vehicle Technologies

## September 16-17, 1992
## Woodfin Suites
## Gaithersburg, Maryland

Sponsored By

Advanced Research
Projects Agency
Software and Intelligent
Systems Technology Office

Edited By

Martin Herman
Robot Systems Division
National Institute of Standards
and Technology
Gaithersburg, MD 20899

## ABSTRACT

The ARPA Unmanned Ground Vehicle (UGV) Demo II program is developing intelligent, semi-autonomous UGVs to perform cooperative tasks in militarily significant scenarios. As part of this program, NIST ran a workshop on UGV performance evaluation in September 1992. The workshop examined the various UGV technologies and aspects of performance that need to be evaluated, including sensing for navigation and driving (vision, stereo, laser, infrared, etc.), planning (mission planning, path planning, etc.), reconnaissance, surveillance, and target acquisition (RSTA), and the integrated perception/planning/control vehicle system. The focus of the workshop was on the breakout of the attendees into working groups. This document presents reports prepared by these working groups.

TABLE OF CONTENTS

# 1. PURPOSE

The purpose of this workshop was to bring together members of the ARPA UGV Demo II team to discuss issues in performance evaluation of technologies relevant to Unmanned Ground Vehicles (UGVs). There were five specific goals for the workshop:

1. To focus the ARPA UGV community on issues of measurement and evaluation of the performance of technologies for UGVs.

2. To emphasize the importance of measurement and evaluation of performance to the ARPA UGV community and to develop a consensus on how to do it.

3. To determine the potential role of each ARPA team member in UGV performance evaluation.

4. To attempt to answer the following questions:
   (1) What can performance measures and evaluation do for the ARPA UGV program?
   (2) Which UGV technologies or aspects of UGV performance need to be measured and evaluated?
   (3) For each of these, what are appropriate parameters, metrics, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?

5. To develop a draft document addressing the questions posed in (4).

## 2. WORKSHOP FORMAT

The focus of the workshop was on the breakout of the attendees into working groups. Four working groups were defined, relating to the following four technology areas:

1. Sensing for navigation and driving
   This group was to consider technologies such as:
   - Stereo
   - LADAR
   - FLIR
   - sonar
   - image segmentation
   - object/landmark recognition
   - road/terrain surface characterization
   - obstacle recognition
   - vehicle acceleration
   - vehicle velocity/position

2. Planning
   This group was to consider technologies such as:
   - mission planning
   - reactive planning
   - path/route planning
   - collision avoidance planning

3. Reconnaissance, surveillance, target acquisition (RSTA)
   This group was to consider technologies such as:
   - tracking
   - target recognition
   - correlation
   - model matching
   - multi-sensor data fusion
   - stabilization

4. Integrated perception/planning/control system
   This group was to consider technologies such as:
   - path tracking
   - road following
   - off-road mobility

obstacle avoidance
coordinated multi-vehicle command and control
operator interface displays and control
communications systems

The purpose of each working group was to develop written recommendations on how to perform evaluations in its respective area. Each group was to consider the issues of

. performance parameters
. measurement techniques
. evaluation criteria
. benchmarks
. test data (content and format)
. test methods
. testbeds

Each group had a chair, whose responsibilites included:

1. Gathering position statements from each of the group members before the workshop.
2. Using these statements as starting point for discussions during the workshop.
3. Focusing the discussions of the group to develop the written recommendations.

The group chairs were the following:

1. Sensing for navigation and driving
   Chairs: Bob Bolles (SRI) & Larry Matthies (JPL)
2. Planning
   Chair: Scott Harmon (Hughes)
3. RSTA
   Chair: Phil Emmerman (ARL)
4. Integrated perception/planning/control system
   Chair: Dave Morgenthaler (Martin-Marrieta)

The workshop started out with just a few talks in the first morning to set the stage by providing requirements of UGV evaluation, i.e., why is it needed and how will it be used to benefit the project. The following speakers made presentations:

Statement of the importance of evaluations for the UGV program
Erik Mettala (ARPA)

Evaluation for the UGV program
Roger Schappell (Martin Marietta)

Performance Evaluation of UGVs
James Albus (NIST)

Testbeds, evaluation and benchmarks for robotic vehicles
Ray Resendes (US Army CSTA)

Image Understanding Evaluation Metrics and Methods
Lynne Gilfillan (LGA, Inc)

In the following sections, reports prepared by the four working groups are presented. A number of position papers were received by each group chair prior to the workshop. All of these position papers are included in their entirety (and only slightly edited) at the end of each working group report.

# 3. BACKGROUND

The success of the ARPA UGV Demo II program is dependent on UGV vehicle and system performance metrics. Performance measures and evaluation can provide quantitative methods for estimating the capabilities and limitations of unmanned ground vehicles under a variety of circumstances. This is crucial to predicting the effectiveness of UGVs under future battlefield conditions, and thus for evaluating the potential benefit of UGVs to the armed forces of the nation. Military planners must understand precisely what the performance characteristics of UGVs really are if they are to integrate UGVs into the plans and requirements documents that are needed for instituting weapons systems procurements. Without DoD requirements documents and procurement plans, UGV research budgets will remain relatively small and unpredictable.

To evaluate the performance of the UGV system or a system component, we can compare with ground truth, with human performance, or with baseline performance determined by a baseline system. Ground truth is the God's-eye view of the scenario, and tells us the precise terrain geometry, terrain material content, positions of landmarks, positions and types of enemy locations, positions and types of targets, etc. If a system component must infer information dealing with some portion of the ground truth, then we can evaluate it's performance in terms of how closely the inferred information corresponds to the ground truth.

A system component's performance can also be evaluated in terms of how closely its output or behavior corresponds to that of a human performing the task under the same conditions.

Finally, the comparison of performance against a baseline system typically indicates how well the advanced versions are doing relative to some previous baseline version.

Although there is no general agreement on the terminology to be used for performance evaluation, the following terms have been used by participants of this workshop.

- Performance parameters, metrics, elements, and dimensions are ways of describing the variables to be measured.

- Performance measures are the methods and techniques for measuring the variables. They require instrumentation for collecting the data.

- Evaluation criteria or evaluation weights are the means by which different variables are combined into an overall evaluation.

- Benchmarks are specific tasks that the system is doing while data are being collected and measures are being done.

## 3.1 Customers and Developers

It was generally agreed by the workshop participants that there are two groups of people who are interested in performance evaluation. The first group consists of the customers, who are interested in evaluating a total product. Examples of customers for UGV systems are Army General Staff, Unmanned Ground Vehicles-Joint Program Office, Office of the Secretary of Defense, ARPA upper management, and Congress. The customers don't particularly care how a certain capability is achieved, or which technology is used. They are primarily interested in overall performance on tasks of interest to them. Performance measures and evaluation are useful to this group of people in that they provide a quantitative measure of how well the system is progressing over time. They also provide a means to determine the potential benefit vs. cost of the system. For example, if a UGV system is to be integrated into a scout platoon, the benefits and costs can be used to determine how much the UGV system will buy in terms of enhanced system performance. In a UGV system under supervisory control, performance evaluation can aid in the allocation of tasks between the human and the machine. Finally, DoD plans and requirements documents can use these measures and parameters to specify their needs.

The second group consists of the technology developers, who develop the algorithms, software, hardware, and integrated systems that constitute a UGV system. The developers want to be able to improve their systems by (1) measuring where are the weaknesses and strengths of the subsystems and (2) measuring whether the subsystem performance improves with certain modifications in the subsystem. Evaluation can help determine when a technique or subsystem is appropriate, characterize its reliability, precision, and limitations. It can also be used to provide focus for future research and development.

# 4. REPORT OF THE WORKING GROUP ON SENSING FOR NAVIGATION AND DRIVING

Working Group Participants:

Bob Bolles (Co-chair), SRI International
Larry Matthies (Co-chair), JPL
Dave Anhalt, Martin Marietta
Mike Daily, Hughes Research Labs
Daniel DeMenthon, University of Maryland
Ramesh Jain, University of Michigan
Keith Nishihara, Teleos
Chuck Thorpe, CMU
Chip Weems, University of Massachusetts
Tom Williams, Amerinex Artificial Intelligence
Al Hanson, University of Massachusetts
Ed Riseman, University of Massachusetts

## 4.1 Evaluating Ability to Sense Terrain Traversability

### 4.1.1 Problem Definition

This problem is loosely defined as estimating those properties of a scene (e.g., geometry, density, moisture content, etc.) important in determining the traversability of a given patch of terrain, with emphasis on off-road navigation. This task has some overlap with other tasks discussed by the sensing group; in particular, with the evaluation of road-following and stereo vision technologies. The discussion here may overlap with discussions of these other tasks.

### 4.1.2 General Comments on Methodology

A common theme in the workshop was that there are several different levels at which performance can be evaluated. This section focuses on metrics to be employed by developers of the technology, rather than on metrics that characterize functionality for the end user. Two contrasting styles of evaluation that were discussed by the sensing group are:

(1) To treat modules as black boxes that take sensor input and produce steering commands as outputs. Evaluation in this case tries to measure the quality of the steering commands as a function of variations in the design of the sensing module. An example of this style of evaluation would be to

evaluate road-followers in terms of smoothness of the steering trajectory.

(2) To distinguish between estimation and control functions within a module and to evaluate each component separately. In the above example, this might involve separate evaluations of how well the road-follower estimated the road center line and how well the embedded control law was able to track the estimated center line.

This distinction applies to determining traversability. The first method combines aspects of sensing and planning into one black box, making it difficult to determine the effectiveness of sensing and planning components. Since the objective of the group was to propose ways to evaluate sensing, the second method above is preferable. It is recognized that there still may be times when the first method is necessary; for example, different approaches to a module might choose different parameterizations of the scene, which would make comparison difficult. The balance of this section will emphasize the second method; that is, evaluation of sensing itself, not bundled into a black box with planning.

### 4.1.3 Characteristics to Evaluate

We identify four categories of scene property that are relevant to determining traversability. These are:

- surface geometry
- surface material type
- density of material
- subsurface structure

The balance of this section elaborates on the meaning of each category and on the role that each plays in the current UGV program.

#### 4.1.3.1 Surface Geometry

This refers to the shape of solid surfaces. This is the most common property estimated in off-road navigation work to date. LADAR and stereo range sensors will be considered as the primary means to estimate geometry.

#### 4.1.3.2 Material Type

This refers to the material of which the terrain or ground cover is made up. Examples include wet or dry soil, rocks, various types of vegetation, and bodies of water. These distinctions are important, because both the size and composition of a terrain feature are important in determining whether the feature can be driven through or must be bypassed. Although exceptions will exist, in general these classes cannot be distinguished with range data alone. Multispectral, polarization, spatial frequency, and other visual characteristics may contribute to estimating material type.

#### 4.1.3.3 Density of Material

Recognizing material type is just a start, because the ultimate need is to determine physical properties of the material, such as density, that bear directly on traversability. Recognition based on visual characteristics can allow inference of such properties based on past experience; however, other types of sensors may allow these properties to be measured more directly. We assume that the UGV program will be largely restricted to visual recognition, so that estimation of density and other properties will be inferred indirectly from visually-based material classification.

#### 4.1.3.4 Subsurface Structure

This is related to density, but the emphasis is on possible existence of subsurface strata with different physical properties that affect traversability. Subsurface voids that might collapse are one example; another is to find the depth of a pool of water and the thickness of soft mud at the bottom of the pool. This category of problem will be considered to be beyond the present scope of the program.

#### 4.1.3.5 Summary

Based on the discussion above, we focus on evaluation for the two categories of surface geometry and material type. To guide the discussion of evaluation, it helps to have a rough model of how these scene properties will be estimated and how they will be used to assess traversability. Therefore, we assume that the primary methods for estimating geometry will be range imaging with LADAR or with stereo vision at visible and/or infrared wavelengths. We assume that material type will be estimated by classification of each pixel, using feature vectors composed of spectral, polarization, spatial frequency, and possibly other

features. We assume that the range data and the labelled images will be combined in some fashion to identify obstacles. As an example, range data may be turned into an elevation grid, cells of which are labelled from the classification data. Obstacle detection algorithms based on surface slope (or other criteria) would be extended to include material type in deciding whether or not a patch of terrain (or location in configuration space) was traversable. Other approaches may be possible; this example is intended to serve as a strawman in the following discussion.

In the balance of this section, we first consider evaluation of surface geometry and material type estimation, then consider evaluation of obstacle detection algorithms built on the lower level sensing capabilities.

### 4.1.4 Evaluation Approaches for Surface Geometry Estimation

The main issue to evaluate is how good are the basic range measurements produced by the sensors; that is, we are not interested here in quality of any surface fitting or other post-processing. Therefore, attention is restricted to evaluating the quality of range imagery produced by (1) LADAR and (2) stereo vision with visible or thermal imagery. If the data format is essentially the same in each case (i.e., an array of range pixels), then essentially the same techniques can be used to evaluate all sensors. It will be assumed that this is true. Below, this section lists metrics that can be used to characterize sensor performance, then discusses what data sets would be useful for such evaluation.

#### 4.1.4.1 Metrics

- Range resolution at each pixel, characterized in terms of standard deviation of the range measurement and spatial covariance of range errors.
    - vs. scene geometry (e.g., angle of incidence is important in characterizing LADAR performance)
    - vs. image noise level
    - vs. image spatial frequency content
    - vs. window size and other algorithm parameters

- Bias in range measurements at each pixel (i.e., error in the mean)

- Spatial resolution
    - in terms of number of pixels across the image

- in terms related to frequency response (i.e., minimum resolvable feature size)

- Sensor-specific metrics for stereo vision, such as percent of pixels with disparity errors exceeding one pixel (i.e., rate of gross correspondence error)

- Computational complexity

- Frame rate with given computing hardware

- Domain of applicability and failure modes

#### 4.1.4.2 Data Sets for Evaluation

A recorded image depends on each scene geometry, spatial intensity function, and image noise; each of these three aspects may be artificial (simulated) or real. Three examples at different points in the spectrum are:

(1) completely synthetic imagery;
(2) real intensity images texture-mapped onto synthetic scene geometry (e.g., simulated ground plane), with synthetic noise added;
(3) real imagery.

It would be desirable to have (1) if it could be achieved with sufficient realism and with low cost. As a practical compromise, (2) is useful for evaluating the bias and precision of stereo range measurements. Item (3) is essential, but the difficulty there lies in obtaining ground truth. A practical compromise in this case is to take images of a set of scenes with simple geometries that cover cases that will arise in more complicated scenes. Since the geometry is simple, ground truth is more feasible to produce. Examples of such scenes are:

- Flat ground
- Flat ground on which simple obstacles are placed; these would include:
  - Vertical range discontinuities (like edge of a rock or wall)
  - Horizontal range discontinuities (like edge of a cliff)
  - Closely spaced pairs of discontinuities, like ruts, potholes, and protruding ridges.

The idea is to characterize performance on simple cases, then extrapolate to complex scenes. Such data sets could be acquired at any site and would not require development of special test courses or data collection facilities. Data sets ideally should include many images for each scene, to allow evaluation of error statistics. To directly compare sensors, data sets of the same scene should be taken with all sensors, under various lighting conditions (day, twilight, night).

With data sets as described above, error statistics would be computed for range imagery produced by each type of sensor, or for various parameters for a given sensor. Evaluation of obstacle detection performance takes place at a higher level; this is discussed later.

### 4.1.5 Evaluation Approaches for Material Type Estimation

Estimation of material type for the sake of traversability analysis can be approached as an application of standard pixel classification techniques, after which other algorithms would use the classification results in assessing traversability. For the sake of evaluation, it seems reasonable to take this as a strawman approach, since it is likely that the ability to accurately label pixels would relate closely to ability to, say, distinguish soil from vegetation by any other means. Thus, the approach (and the evaluation) break down into two phases: (1) estimating the material (i.e., pixel classification), then (2) reasoning about traversability, given the material (and geometry).

The problem of evaluating classification accuracy has been addressed for a long time for remote sensing purposes. In a nutshell, it involves establishing ground truth classifications by manual means, then measuring error rates as a function of feature set and classification algorithm. Much the same methods are appropriate for UGV applications.

Reasoning about traversability, given combined information about geometry and material, again is a higher level function that will be addressed next.

### 4.1.6 Evaluation of Obstacle Detection Performance

The basic process model being addressed here has the stages

sensing (range, material type) => obstacle detection => path planning.

Previously we addressed evaluation of the sensing stage. Obstacle detection

builds on sensing to detect obstacles and record their locations, either relative to a map of the terrain or relative to a map of vehicle configurations (i.e., configuration-space). Path planning takes the obstacle representations as input and produces desired vehicle trajectories as output. Obstacle detection and path planning may be interleaved in a lazy evaluation type of search, but the obstacle detection component will still be recognizable. For the purpose of this report, we include obstacle detection within the "sensing" function and propose methods to evaluate its performance.

The issues in evaluating obstacle detection are (1) whether or not obstacles are detected and (2) how accurately they are localized. Detection breaks down further into two cases: correct detection of obstacles that are present, versus incorrect "detection" when there are no obstacles present (false alarms). Detection and false alarm rates, as well as localization, should be characterized as a function of (1) scene characteristics (e.g., size and distance of the obstacle), (2) sensor characteristics (e.g., LADAR vs daylight stereo vs thermal stereo), and (3) algorithm characteristics (e.g., image resolution, details of the obstacle detection algorithm). Ideally, data sets and evaluation methodologies can be defined that are independent of these characteristics.

Consider first evaluating false alarm rates. If a flat road is considered to have no obstacles, then false alarm rates can be assessed simply by driving down the road and recording perceived obstacles, for example as average number per unit distance per frame. Recording such imagery for a variety of road types would provide a good ensemble of different "scenes" to indicate average false alarm performance.

False alarms can also occur when small bumps are present that are below the threshold of what is considered an obstacle that must be avoided. This case is a little more difficult, because it is harder to establish ground truth for bumpy roads than for smooth roads. Nevertheless, a basically smooth road can be sprinkled with objects of known sizes to test performance with bumps that are below the true obstacle threshold.

Rates of correct detection can be evaluated in a similar fashion. To simplify logistics, one might use (say) a wide dirt road, drive the vehicle down the right side of the road, sprinkle obstacles along the left side of the road, and bias the vision system to be left-looking. All obstacle detections would be recorded and compared with ground truth information about the obstacle locations and sizes; the result would be to characterize probability of detection as a function of

obstacle size and distance. It may be difficult to register the imagery with ground truth obstacle data; this could be facilitated by placing high-contrast place markers along the road for automatic logging of distance along the road via image processing.

Another approach to evaluating rates of correct detection is to leave the vehicle stationary and to place obstacles of known sizes at known distances from the vehicle. This is much simpler (and has been done already to some degree at JPL), but it provides a much more limited ensemble of background scenes than can be obtained by the driving approach described above.

In creating test scenes for evaluating detection rates, object material types can be intermixed to evaluate this variable. For example, some "obstacles" may be rocks and others may be bushes of similar size to the rocks. This can be done for data sets taken with stationary or moving vehicles.

### 4.1.7 Summary of Recommendations

1. Consider scene geometry and material type in evaluating traversability.

2. Distinguish three stages of processing (sensing, obstacle detection, and path planning) and conduct separate evaluations of the sensing and obstacle detection stages (path planning is the province of another working group).

3. For sensing scene geometry, evaluate quality of data produced by imaging range sensors (LADAR and stereo with daylight and thermal cameras):
   a. Use metrics designed to measure error statistics.
   b. Collect data sets that allow measurement of these error statistics.
   c. Collect side-by-side data sets wherein all sensors view the same scene under various lighting conditions (day/twilight/night) to allow comparison of performance.

4. For sensing material type, assuming that pixel classification is part of the process, apply customary techniques involving manual specification of ground truth, followed by measurement of error rates as a function of algorithm design.

5. For obstacle detection:
   a. Characterize probability of detection and probability of false alarms as a function of:

- scene characteristics (e.g., obstacle size and distance)
- sensor characteristics (e.g., type of camera, level of noise)
- algorithm characteristics (e.g., image resolution, details of obstacle detection algorithm)

b. Use the following data sets:
- stationary vehicle observing known obstacles at various distances (evaluates detection for a small number of scene backgrounds)
- moving vehicle on flat roads (evaluates false alarm rates for a large number of scene backgrounds)
- moving vehicle on flat roads with known obstacles (evaluates detection rates for a large number of scene backgrounds, but requires more effort in data collection and ground truth preparation)

c. The above does not include data sets for rough, off-road terrain. It will be more difficult to establish ground truth for such cases. The argument was made that evaluating performance under simpler geometries should suffice, because rougher terrain will just include many instances of the simpler geometries. Even if this is not true in the long run, it represents a practical starting point.

## 4.2 Stereo Sensor Evaluation

### 4.2.1 Metrics and measures

#### 4.2.1.1. Camera noise tolerance

This deals with the signal to noise level at which the matcher performance break down. There are various kinds of noise that must be dealt with by stereo matchers. These include (white) shot noise from the sensor, coarser pattern noise from the sensor arrays, and quantum noise when working with high gain cameras. There are also other effects such as eye position dependent shading, shadows, and occlusion that might be considered a noise effect at the stereo matcher.

Even in bright daylight scenes, shot noise and pattern noise can be significant factors affecting performance. For example, the very low contrast texture present over something like a uniformly painted wall would be dominated by camera noise sources.

In support of the UGV effort, it would be valuable to have numbers that

characterize a given algorithm's noise thresholds. It would be useful to define a suite of tests that measure things like the noise level where matching performance drops below some level. Performance might be a combination of probability of making a match and standard deviation of those matches from ground truth.

This can be effectively measured using a suite of test stereo pairs of a calibrated scene including textured flat surfaces as well as various sized objects with increasing noise levels. An effective way to increase the noise level is to reduce the scene lighting in increments down to the point where the noise dominates the image by an order of magnitude or so. Setting things up so that disparity varies linearly across an extended area would make it easy to compute measurement standard deviations.

#### 4.2.1.2. Disparity gradient limit

This deals with the orientation of the surface being imaged. How much can the flat surface slope away from the image plane before matching breaks down?

There is a disparity gradient effect that is likely to be a significant factor limiting matcher performance on planar ground surfaces as imaged from the UGV. The magnitude of the disparity gradient is approximately related to the ratio of camera baseline and camera elevation. This restricts the use of larger camera baselines by the disparity gradient tolerance of a given matcher.

A natural way to measure this performance parameter is with a suite of stereo images made from a flat textured surfaced imaged to produce a series of different disparity gradients.

#### 4.2.1.3. Disparity resolution-speed product

In addition to wanting to quantify an algorithm's matching reliability, we need to measure the speed at which the matching can be accomplished in a UGV context. Various algorithms search different disparity ranges, they produce various degrees of subpixel resolution, and they cope with different amounts of vertical disparity uncertainty. The following expression takes these differences into account:

(disparity range searched)/(standard deviation of measurements)
*(vertical disparity capture range)*(measurements per second)

There are a variety of ways that the "measurements per second" term could be derived. For example, on a standard CPU platform, or on whatever platform is most appropriate to the computation provided, it is compatible with the UGV architecture.

#### 4.2.1.4. Disparity measurement latency

This deals with how long it takes from photons arriving in the camera lens to measurement out of the system.

For realtime systems, it is also important to know how long it takes to get any data back from the sensor system. A system that produces dense range arrays, for example, may have a very high resolution-speed product, but it may take several seconds to yield any of that data to subsequent analysis.

#### 4.2.1.5. Spatial resolution

a. How small an object is detectable (diameter in pixels) assuming it is within the disparity search range.

b. How well can an extended depth edge be located?

c. How well can range to a smoothly undulating surface be measured?

#### 4.2.1.6. Range from disparity precision

In addition to matching algorithms, we must deal with issues of calibrating our sensor systems and computing range from disparity efficiently. Different approaches to calibrating and computing range from disparity should be compared with metrics like:

a. Complexity of calibration procedure; how long does it take; does it require an elaborate array of calibrated targets? How easy is it to make a field change such as changing lenses or remounting cameras?

b. Accuracy of computed range measurements from disparity plus vergence measurements.

c. How long does it take to compute range from disparity?

### 4.2.2 Techniques

Looking beyond the generic performance metrics listed above, we also need to look more specifically at the UGV mission requirements and relate those metrics to it. In addition, it will often be expedient and a powerful sanity check to run prototype algorithms through realistic mission scenarios where high fidelity recorded stereo imagery along with all other pertinent vehicle data are played back at realtime rates.

### 4.2.3 Test data, testbeds, and benchmarks

#### 4.2.3.1 Calibrated testbenches

Where range, surface slope, surface material, lighting and sensor noise can be controlled would be very useful for making many of the above measurements. Image data sets, both digital data and continuous video, collected from such testbenches could be distributed to all participating researches to provide a standard means for comparing and evaluating performance.

#### 4.2.3.2 Field data

High quality continuous video recordings from the actual test ranges together with high performance navigation, vehicle and sensor attitude data will be critical to the development and verification of stereo algorithms for deployment on the UGV. Further annotation on the data stream of actual hazards present in the opinion of the driver during the test run would be extremely valuable.

## 4.3 Evaluating Landmark Recognition

Landmark recognition consists of three primary phases

1) Recognition or detection in the case where a model already exists for a landmark;

2) Tracking phase over a sequence of frames or snapshots;

3) Landmark Acquisition - builds a model of a landmark to use in the other two phases.

Both start-up and continuation can have associated metrics and evaluation. We did

not concentrate on the problem of evaluating the process of creating new, on-the-fly landmarks, but rather, we concentrated on detection of known landmarks, either (as in 1 above) not yet sighted, or (as in 2 above) previously sighted.

As with many evaluations, one needs to set up measures, and then use data which span all the reasonable conditions which will tax the system. In the case of landmark recognition, there are a number of identifiable conditions, resulting in a coarsely quantized, high dimensional space of conditions, and only three important measures.

### 4.3.1 Measures

The primary performance criterion is:

1) Accuracy of derived vehicle position (in terms of distance and pose from actual position - this is a six degree-of-freedom measure).

Secondary (and more useful to the development of the technology) criteria are:

2) Accuracy in pixels and radians of the direction of a known landmark

3) Correctness in finding the given landmark
Three conditions are useful in this measure, the first two are errors, the third is "correct":
    1) if the landmark is in view and not found,
    2) not in view and found,
    3) in view and found.

### 4.3.2 Conditions of variation

Conditions under which the evaluation criteria are to be applied can be used to define a data set. The programs to be tested would be applied to the data set, and the measures applied to the program results, and statistics of the measures would characterize the evaluation.

1) Varying lighting conditions (day, night, morning)

2) Varying weather (snow, rain, wet, dry, humid, cold, hot)

3) Varying season (foliage dead, in bloom, fruited)

4) Varying sensor noise (electronic noise, jitter, blur)

## 4.4 Individual Position Papers

### 4.4.1 Bob Bolles, SRI

WHAT CAN EVALUATION DO FOR THE UGV PROGRAM?

On the upside, evaluation could help determine when a technique is appropriate, characterize its reliability and precision, and document its quirks, all of which would help a person apply the technique appropriately and point out problems for future research. Over time, a sequence of evaluations could show the progress in the development of techniques for a specific task.

On the downside, evaluation could take over a subfield, drain it of scientific vitality, create discord within the community, and lead to dramatic funding cuts, all of which could drive a researcher to drink.

My conclusion at this time, although I have some mixed feelings about it, is that the time is ripe to start a constructive thread of evaluation within computer vision, keeping in mind that we're new at this, so we shouldn't take it too seriously too soon. As long as we keep it as one of the many facets within the UGV program and we emphasize cooperation, openness, and scientific curiosity, then I think it can be a constructive tool.

WHICH THINGS SHOULD BE EVALUATED?

Every group, whether it is a primitive component (such as stereo and LADAR) or a whole subsystem (such as road following and navigation), should develop its own thread of evaluation within their community.

HOW SHOULD TECHNIQUES BE EVALUATED?

My comments are based primarily on my limited experience with the evaluation of stereo techniques. My first general observation is that it takes a lot of work to do an evaluation--you have to determine the scope of it, gather a data set, homogenize the data set (i.e., put it in a shareable format), document it, distribute it to participants with instructions, gather the results, homogenize them, write programs to analyze the results, write up the statistics, make general

observations, write up the observations, distribute the results, statistics, and observations for comments, and plan for the next phase.

Unfortunately, our first attempt at stereo evaluation has taken so long that our observations will not be quite as timely as they should be. Hopefully, the second phase will be significantly easier since we have (1) established communication formats, (2) each participant has implemented an end-to-end system for analyzing the data and producing results in the appropriate formats, and (3) we have implemented techniques for examining the results and gathering key statistics.

My second observation is that each evaluation has a limited scope, and hence should not be used to make sweeping conclusions. Or put another way, all the stereo techniques we've looked at have their strengths and could be rated "the best" in some reasonable evaluations. Therefore, it does not make sense to say that X is the best technique, except under a specific set of conditions. Plus, real tasks typically include many more "conditions" than can be covered by any one of the current techniques.

Third, ground truth is crucial, but hard to come by. In our initial stereo test we only had a limited amount of ground truth, which meant that we had to compare the results of one technique to another or to interactively measured values. This is unsatisfactory because the results are not being compared to the "real" values and because it is expensive to measure even a few points interactively.

Fourth, gathering a data set inherently limits the type of techniques that can be applied. For example, conventional stereo data is not sufficient for trinocular techniques. Therefore, even if a trinocular technique may be better, there is no way to include it in the evaluation and compare results without gathering trinocular data. Another example: gathering data from a moving vehicle without annotating the data with motion information significantly restricts the types of techniques that can be applied.

Fifth, and this is not an observation, but rather a gut feeling, it is extremely important to stress openness, sharing, and constructiveness. Without these, everything can fall apart in a hurry. The stereo techniques that we included in our evaluation were developed over many years, which means that their developers have their egos tied up in them, some more than others. In addition, since these techniques were designed and developed for specific domains

(such as cartography) or within a certain paradigm (such as active vision), it is not surprising that their behavior may not be as good as their developer would like on our test data. On the other hand, I am happy to say that all the techniques did pretty well. They found reasonable matches for most of the points in the images and avoided reporting false matches when there weren't any. Of course, all the techniques had problems, including producing a few gross errors. But their worst problem was that they left "holes" in their results where matches were possible. I expect that all of them could do quite a bit better with a little more work (primarily in the development of heuristics for setting different parameters for different parts of an image). We'll see about this in our next phase.

As part of our policy to be open and constructive we plan to distribute all the results to all the participants. In addition, we will supply a detailed description of each group's approach (written by the group), a copy of each group's analysis of their results, and a copy of our analysis. We plan to annotate our statistics as thoroughly as possible to indicate strengths of each technique, and when possible, to hypothesize why a technique made a mistake.

## TECHNIQUES

For our stereo evaluation we (JPL, SRI, and Teleos) have informally adopted a three-pronged approach to evaluation:

(1) Analytic equations to model geometric relationships and answer such questions as the expected range precision possible from a particular camera configuration.

(2) Qualitative analysis to locate technical voids, such as problems with shadows, horizontal edges, or repeated patterns.

(3) Statistical analysis to estimate the expected precision of a technique and such things as the probability of detecting an object at a specified distance.

Unfortunately, although all of us have equations representing stereo configurations, no one has taken the bull by the horns and developed, documented, and distributed a complete analytic model. This should be remedied in the near future.

We at SRI have spearheaded the qualitative analysis and are close to finishing the first phase of our on-going evaluation process. JPL has undertaken the statistical analysis, and they too are finishing up their first pass.

## METRICS

The basic metrics of reliability, precision, and speed are sufficient for most sensor evaluation. For our "qualitative" stereo evaluation at SRI, we have concentrated on reliability for two reasons. First, we didn't have enough ground truth information to make meaningful precision estimates. And second, the current implementations of the techniques were on such machines as SUNs and LispMachines. In the future we expect to gather more ground truth and test implementations for which speed is a serious consideration.

For our initial tests we worked with the following measures of reliability:

- Percentage of matches (correct or not) returned in matchable regions.
- Percentage of incorrect matches in matchable regions (false positives, type 1) -- generally, without ground truth, we had to estimate these by hand.
- Percentage of (incorrect) matches returned in unmatchable areas, such as occluded regions and regions out of the field of view of one of the images (false positives, type 2).
- Percentage of matchable points for which no match was returned (false negatives).

In addition to making these measurements over the whole image, we also gathered them in special (hand-drawn) regions of the images covering such things as foliage, ground, tree trunks, and shadowed regions. We did this to identify types of regions in which the techniques were particularly weak.

In addition to these raw statistics, we are also listing the quirks of each techniques. An example of a quirk is an odd pattern of results caused by a particular search strategy. Another example is a degradation of results due to inaccurate calibration.

DATA AND TESTBEDS

As I've already mentioned, key parts of any data set are the ground truth of the sensed scene and the auxiliary data associated with data acquisition, including such things as the sensor configuration, internal sensor parameters, and the vehicle motion parameters. These are expensive to measure, but crucial to evaluation.

Clearly, the more realistic the data is, the better. In other words, gathering data from the demo vehicle at the demo site is the best (assuming that you gather ground truth and auxiliary data). Gathering data from a similar sensor configuration mounted on a similar vehicle at another, possibly instrumented, site is almost as good. When you get further from the demonstration area and vehicle, the relevance of the evaluation drops significantly.

One other possibility is to generate synthetic data, which offers completely known auxiliary information and ground truth. Unfortunately, it has been difficult to generate realistic-looking and sufficiently complex data sets to be of much interest. Recently, however, at SRI we have developed a technique to modify real stereo pairs based on our stereo analysis to generate significantly better, although still not perfect, synthetic stereo pairs. We plan to incorporate a few examples of this data in our next data set.

PLANS FOR PROGRAM

Since real data from the demonstration sites is the best, I think it is important for Martin Marietta to have the equipment to perform all three tasks required for gathering complete evaluation data sets: recording multiple live cameras, computing and recording auxiliary data, and measuring ground truth. In addition, there should (could) be other sites, such as a testbed at NIST that has significantly more complete ground truth or a site at JPL where the stereo group and Hughes could run somewhat more limited experiments.

### 4.4.2 Larry Matthies, JPL

*What can performance measures and evaluation do for the ARPA UGV program?*

(1) Quantify:

- Performance requirements of integrated system.

- Performance requirements of component technologies, based on impact on overall system performance.
- Performance achieved by component technologies, in terms of quality and speed versus cost.

(2) Provide focus for future research.

*Which UGV technologies need to be measured and evaluated?*

I'll restrict my comments to range sensors and to processing applied to range data. For sensors, we need to evaluate LADAR, stereo with daylight, intensified, and FLIR cameras, and other acoustic or low-frequency radar for skirt sensing. For processing, we need to evaluate reliability of obstacle detection, landmark recognition, road-finding, etc. as a function of the quality of the input range data. Finally, all of this needs to be contrasted against the cost of the achieved performance.

*What are appropriate metrics, test data, etc?*

(1) Metrics

We need metrics for the quality of the range data, the quality of properties estimated from the range data, and the cost of the estimation procedures.

Metrics for quality of the range data:

- Bias:
  - Mean squared error
- Dispersion:
  - Variance
  - Entropy (for stereo in particular)
- Reliability of confidence measures
  - Mean and variance
- Spatial resolution (not just in terms of pixels, but also in terms of frequency response)

Metrics for quality of derived properties:

- e.g., obstacle detectability:
  - Probability of detection and false alarm

Metrics for cost:

- Dollar cost of the hardware
- Computational complexity
- Number of bits/pixel, size/number of stages of convolution and correlation
- Computing time on given hardware

(2) Test data

Test data will be most valuable if it comes with good calibration data. Therefore, evaluation efforts should address the calibration issue early on. This suggests that the research community should attempt to standardize on camera models, calibration procedures, and metrics for modeling/evaluating the quality of calibration. This will be valuable even for algorithms that can cope with certain lack of calibration, because this will allow us to be explicit about what is gained or lost thereby.

I advocate the collection of large samples of images to allow estimation of sample statistics for the results. Examples include (1) multiple images of the same scene and (2) multiple images with the same 3-D geometry, but different intensity distributions frame-to-frame. An example of the latter is images taken at intervals while driving along a flat road. It is also desirable to acquire side-by-side imagery of the same scene from all of the candidate range imaging sensors, including LADAR, visible light cameras, FLIR, and image intensifiers. Such data should also be taken at different times of the day (from identically the same viewing position to allow direct comparison), such as morning, afternoon, twilight, and night.

Data sets like those proposed above will be massive. This suggests that we address the question of how to archive and distribute the data; for example, is it time to go to CD-ROM or optical disk?

As usual, it is desirable to have ground truth information about the 3-D geometry, the uncorrupted intensity image, and the noise distribution, but this information tends to be hard to come by. Therefore, real and synthetic images, scenes, and noise fields all have a role to play.

Finally, data sets are desirable for scenes with and without water and various types of vegetation, as well as for different camera configuration (e.g., number and placement of cameras).

### 4.4.3 Mike Daily, Hughes Research Labs

(Scott Toborg, Teresa Silberberg, and Kurt Reiser contributed to these comments.)

Four critical needs that must be addressed in defining meaningful performance evaluation:

1) The need for modular systems with well-defined interfaces to sensors, hardware, and other procedures. The most important method for evaluating performance should be how the vision module performs in the context of the successively broader system, including the overall system (i.e., the UGV). Vision system modules should be designed so they can be extracted and replaced by new modules. This allows the resulting improvement or degradation in overall system performance and capability to be compared over many approaches.

2) The need for common hardware, languages, and user interface software. Comparison across different hardware platforms including parallel and pipelined architectures is difficult, especially for timing. Reimplementation in different languages often changes results (i.e. how a float is defined). Use interface code should be separate from the important implementation details so rehosting is less time consuming.

3) The need for common databases of imagery for different sensor types and for static and dynamic algorithms. Testing algorithms and vision modules on common data is imperative. Two types of data are needed: static and dynamic. Static data is useful for modules that do not interact with the sensors or world. This includes many object and landmark recognition methods. Dynamic data is needed for modules that control the sensor actively or that interact with the world (perhaps through the vehicle). In both cases, ground truth is important for verification. Static image data (which may also include motion sequences) for different sensor types and a wide range of environments representative of what is expected in applications such as UGV would be very useful in aiding performance evaluation (a common image database). Dynamic databases are more difficult to provide since sensor or vehicle control influence the position of the sensors. A validated simulation of sensor data would provide the best means to evaluate algorithms that need to acquire new images

interactively. Validation of the simulation relative to the real world will ensure that simulated images are adequate for algorithm evaluation and potentially for algorithm development. We are using such a simulation (Realscene) currently to perform integration of vehicle control behaviors and ALVINN for road following.

4) The need for standard, accepted methods of degrading or stressing algorithms and vision modules. We need to be able to determine the range of inputs that a given module can use. Simply adding noise to images may not be good enough. This is partly dependent on obtaining representative databases that span the space of important environments. Vision modules should be evaluated on hundreds or even thousands of images that will stress the system to show where it fails.

Other issues:

- Who will do the evaluation?
- Is the appropriate evaluation method significantly different for different sensors, environments, applications?
- Is evaluation of the parts of a vision system adequate to ensure the performance of the whole system?

Haralick (in Jan. 92 Pattern Recog. Letters) proposes this approach (roughly):

1) Generate representative images (simulated or real).
2) Randomly perturb and add noise.
3) Characterize performance using some pre-defined criterion function measuring the difference between "ideal" unperturbed data and the perturbed input.
4) Analyze data over many experiments.

This approach is adequate for some algorithms, but doesn't seem realistic for lots of others. It also requires the definition of an ideal input, which may be too difficult in outdoor domains.

### 4.4.4 Chip Weems, University of Massachusetts

The basic point is that any benchmark has a technical element and a political element. The technical element is the part that must be gotten right in order to get useful measurements from the benchmark. The political element is the part that ensures participation in the exercise and acceptance of the results. The political

part is actually more important in some ways to the success of the exercise.

ARPA has to force participation with workshops and hard deadlines for results. The benchmark has to be designed as a collaborative effort to diffuse blame -- the design team should include representatives of different groups, and should actively solicit comments from the whole community.

Technically, realism is the key to the getting usable results, but it can also make the task too difficult to carry out. Thus, a balance must be struck. Benchmarks can be too strongly or too weakly constrained --the former stifles creativity and innovation, the latter can result in uncomparable data. The combination of strong and weak constraint (e.g., solve the problem this way, then solve it again any way you like) is one approach.

Any evaluation should be structured to avoid a horse-race mentality, otherwise there is too great an incentive to cheat, and the results become worthless. It helps, for example, to only make direct comparisons between present and past performance, or between performance of individual modules within a larger task. Having any kind of bottom-line figure of merit will hopelessly skew the results.

Remember: Benchmarks are a class 5 mendacity, falling between Damn Statistics and Delivery Promises on the ANSI standard mendacity scale.

### 4.4.5 Keith Nishihara, Teleos

Introduction
- Statement of the purpose of the report
- Evaluation's potential contribution to the UGV Program
- Technologies to be evaluated (LADAR, landmark recognition ...)
- Description of the process taken to write this report

I would suggest including the two sections below along with a discussion of relevant performance metrics. Following sections on specific sensing technologies can point back at these sections with notes on where they can be applied and how to evaluate them to satisfy the mission task metrics.

KEY MISSION TASKS
- on road following
  - speed, types of road, weather and road conditions
- off road obstacle avoidance

characterization of types of terrain, weather conditions
characterization of dangerous land features
sizes, shapes, other qualities such as coloration, reflectivity, etc. that can affect their detectability
necessary and desired speed ranges
designation of test areas for data collection and performance evaluation
navigation to destination
how much use of what we sense along the way must/can be used
return navigation
others....

KEY MISSION CONSTRAINTS
cost, availability, and bulk limitations on sensor systems
availability and accuracy of onboard position and attitude information
detectability (passive to active spectrum)
robustness of sensors/algorithms/hardware/and code
(how much tailoring to specific terrain will be allowed)
others....

OPTICAL SENSORS (it will be useful to have these characterized under UGV mission conditions)
CCD (bw and color),Image intensifiers, IR etc
noise characteristics, different day light conditions, night, etc
spatial resolution, temporal resolution, failure modes (eg blooming)

LADAR
metrics & measures
techniques
test data, testbeds, & benchmarks
plans for the program
Landmark recognition
metrics & measures
techniques
test data, testbeds, & benchmarks
plans for the program
Obstacle recognition
. . . .
Stereo

. . . .

Surface characterization

. . . .

Conclusion

STEREO

A. Metrics and measures.

A.1. Camera noise tolerance (at what signal to noise level does matcher performance break down)

There are various kinds of noise that must be dealt with by stereo matchers. These include (white) shot noise from the sensor, coarser pattern noise from the sensor arrays, and quantum noise when working with high gain cameras. There are also other effects such as eye position dependent shading, shadows, and occlusion that might be considered a noise effect at the stereo matcher.

Even in bright daylight scenes, shot noise and pattern noise can be significant factors affecting performance. For example, the very low contrast texture present over something like a uniformly painted wall would be dominated by camera noise sources.

In support of the UGV effort, it would be valuable to have numbers that characterize a given algorithm's noise thresholds. It would be useful to define a suite of tests that measure things like the noise level where matching performance drops below some level. Performance might be a combination of probability of making a match and standard deviation of those matches from ground truth.

This can be effectively measured using a suite of test stereo pairs of a calibrated scene including textured flat surfaces as well as various sized objects with increasing noise levels. An effective way to increase the noise level is to reduce the scene lighting in increments down to the point where the noise dominates the image by an order of magnitude or so. Setting things up so that disparity varies linearly across an extended area would make it easy to compute measurement standard deviations.

A.2. Disparity gradient limit (how much can a flat surface slope away from the image plane before matching breaks down?)

As discussed at the previous UGV workshop, there is a disparity gradient effect that is likely to be a significant factor limiting matcher performance on planar ground surfaces as imaged from the UGV. The magnitude of the disparity gradient is approximately related to the ratio of camera baseline and camera elevation. This restricts the use of larger camera baselines by the disparity gradient tolerance of a given matcher.

A natural way to measure this performance parameter is with a suite of stereo images made from a flat textured surfaced imaged to produce a series of different disparity gradients.

A.3. Disparity resolution-speed product

In addition to wanting to quantify an algorithm's matching reliability, we need to measure the speed at which the matching can be accomplished in a UGV context. Various algorithms search different disparity ranges, they produced various degrees of subpixel resolution, and they cope with different amounts of vertical disparity uncertainty. The following expression takes these differences into account:

(disparity range searched)/(standard deviation of measurements)
*(vertical disparity capture range)*(measurements per second)

There are a variety of ways that the "measurements per second" term could be derived. For example, on a standard CPU platform, or on whatever platform is most appropriate to the computation provided it is compatible with the UGV architecture.

A.4. Disparity measurement latency (how long from photons in lens to measurement out)

For realtime systems, it is also important to know how long it takes to get any data back from the sensor system. A system that produces dense range arrays, for example, may have a very high resolution-speed product, but it may take several seconds to yield any of that data to subsequent analysis.

A.5. Spatial resolution

A.5.a. how small an object is detectable (diameter in pixels)
assuming it's within the disparity search range

A.5.b. how well can an extended depth edge be located

A.5.c. how well can range to a smoothly undulating surface be measured.

A.6. Range from disparity precision

In addition to matching algorithms, we must deal with issues of calibrating our sensor systems and computing range from disparity efficiently. Different approaches to calibrating and computing range from disparity should be compared with metrics like:

A.6.a. complexity of calibration procedure, how long does it take, does it require an elaborate array of calibrated targets? How easy is it to make a field change such as changing lenses or remounting cameras?

A.6.b. accuracy of computed range measurements from disparity plus vergence measurements.

A.6.c. how long does it take to compute range from disparity?

B. Techniques (I presume this refers to methods for obtaining above performance measures which has already been covered to some degree above)

Looking beyond the generic performance metrics listed above, we also need to look more specifically at the UGV mission requirements and relate those metrics to it. In addition, it will often be expedient and a powerful sanity check to run prototype algorithms through realistic mission scenarios where high fidelity recorded stereo imagery along with all other pertinent vehicle data are played back at realtime rates.

C. Test data, testbeds, and benchmarks

C.1 Calibrated testbenches

Where range, surface slope, surface material, lighting and sensor noise can be controlled would be very useful for making many of the above measurements. Image data sets, both digital data and continuous video, collected from such testbenches could be distributed to all participating researches to provide a standard means for comparing and evaluating performance.

C.2 Field data

High quality continuous video recordings from the actual test ranges together with high performance navigation, vehicle and sensor attitude data will be critical to the development and verification of stereo algorithms for deployment on the UGV. Further annotation on the data stream of actual hazards present in the opinion of the driver during the test run would be extremely valuable.

D. Plans for the program

(needs to be discussed at the meeting)

### 4.4.6 Tom Williams, Amerinex

I. GENERAL COMMENTS ON EVALUATION OF I.U.

A. AUDIENCE

I have interest in metrics for IU, and will be presenting most of these ideas at the ARPA IU PI meeting in Utah. I strongly believe that there are different metrics for different audiences, and that one of the objectives of ARPA is the "selling" of research to Congress and DoD. Metrics for this purpose are very different from those intended to assist researchers in determining whether one approach is "better" than another, or more likely, whether a particular algorithm has "improved", or works at a certain performance level on more data. Therefore, unless you want to press on with the second reason for evaluation - ARPA interest in measuring progress for the purpose of convincing others - I will contain my comments to the measurement of what I call "fundamental IU technologies". This measurement roughly translates into classes of algorithms, or what you call "characterizing the effectiveness of techniques" rather than the performance of systems. System performance is an outward behavior of a complex of interacting algorithms/techniques. It is system performance which ARPA can use to convince a congressman of progress, e.g. it drives through the field and shoots in the dark, whereas algorithm effectiveness is only of interest to scientists, e.g., 73% of the vector field is within 1% of correct depth, prior to smoothing.

With all this said, I believe that we can make some statements about whom the audience could be for different types of metrics. These statements should be in

the Introduction. It has bearing on the questions:

- What can we realistically expect from evaluation?
- How can we use evaluation in as constructive a way as possible?

These differences in purpose/audience also have impact on encoding the results, obviously for planners and scientists.

## B. ABUSE

Unfortunately, such results can also have a devastatingly negative impact on the scientific process. It can focus people on certain paths which seem more promising and reduce the likelihood that researchers will try new and weird ideas, or even pursue slightly tangential areas or mix disciplines.

Consider an analogy, the IQ test. It is a good analogy since we are effectively trying to measure an aspect of machine intelligence (it may well be a poor analogy for other reasons). Our first observation is that an IQ test is a very difficult thing to create. The problem is a compound of 1) attempting to measure things that are not easy to test - we can only observe the thinking process by answers to questions, and 2) attempting to be fair by covering as many aspects of the human intellect as possible. The fairness problem arises because many assumptions have to be made, whether or not we are aware that we are making them.

Ultimately, the real test of intelligence is the success of the being as it interacts with the world. The IQ test, however, measures performance of humans in vocabulary tasks and visual manipulation, rather than measuring a human's value to society. Similarly, we might create a measure of performance of machines in stereo depth mapping or obstacle avoidance, and possibly miss the real value of autonomous machines to the Army.

One final consideration of this analogy, is that IQ tests can be used to see if an individual's intelligence changes, and if certain educational techniques are effective. It also can be abused, to determine who gets a job, who is superior, and worse, teachers can teach "to the test" stressing only those skills that are tested. Given the imperfect nature of the test, this is ultimately detrimental to the individual and the society. Similarly, these tests for IU could be used to determine if an algorithm's performance improves, or if one technique is more appropriate than another, and can also be abused, to determine which research group gets funded, which institute is superior, and worse, implementors can

specialize algorithms to perform well on tests - like compilers are optimized for Linpacks - leading to specialized solutions and no theoretical advances. We need to have safeguards against abuse; the community acceptance of metrics will depend on such safeguards.

## II. INPUTS TO THE WORKSHOP SENSOR GROUP OUTLINE

### A. INTRO

Evaluation's potential contribution to the program.
Evaluation will provide a means for determining, given a set of circumstances --- i.e., a particular data set, sensing conditions, (atmospheric, time of day, glint, camouflage) the following:

1) whether a specific algorithm has, through modification, improved
2) whether a specific algorithm is better than another
3) that a specific algorithm performs at a certain level
4) that certain data lacks fidelity
5) that certain data has questionable ground truth

Technologies to be evaluated:

I think your list is a good start, but is not specific enough, you are mixing techniques with systems. My list would be only techniques:

- Active range based obstacle detection
- Stereo techniques for dense range map extraction/surface reconstruction
- Stereo techniques for model based object recognition
- Stereo techniques for sparse range map extraction
- Stereo techniques for obstacle detection
- Passive imaging techniques for landmark recognition/pose refinement
- Passive imaging techniques for object recognition
- Passive monocular techniques for temporal based (various) range mapping
- Passive monocular techniques for temporal based object recognition
- etc. etc. etc.

My argument for this level of detail, is that 1) one cannot compare apples to

oranges, 2) we won't really know if one algorithm is better than another at solving a problem unless we know just what problem it is trying to solve, 3) different institutions are really taking significantly different approaches to solving problems, and we should be careful to understand the common ground and differences.

B. MEASURES

There are a variety of measures. In fact, it is quite hard to decide on measures which are meaningful across a wide variety of algorithms, so I would argue for different measures for different techniques (classes of algorithms). The following is the list I will be presenting at the ARPA IU PI meeting:

Measures would consist of a combination of measurements which compare the results of running an algorithm on a specific data set with a ground truth set for that data. There are three basic measurements that I propose, each with various qualifications.

1) Accuracy (generally correctness and precision). This usually means correctness of description. One either does or does not have the correct description for each object (or vector, or depth point, etc). This implies that the application is defined, and that all objects on which the application depend have associated truth. For each technique there may easily be other accuracy measures, such as correctness of depth, surface angle, and size. Here the need for excruciatingly accurate ground truth is much greater, and the collection task more difficult. Indeed, how do we represent image and scene truth? Accuracy also implies a variety of individual measures. Typical questions that must be answered are; how does one combine individual errors, is there some single measure, how does one report false negatives and false positives, at what level of abstraction should measures be taken (lines, regions, surfaces, objects, groups of objects)?

2) Coverage refers to the completeness of the description. If the application requires a dense depth map, then the density of the results of an algorithm will be important in its metric. The choice of level of abstraction of image and scene events usually determines the number of items that will participate in a "coverage" metric. The level of abstraction is somewhat tied to the technique, but most techniques can have several abstraction levels important for measurement.

3) Speed is perhaps the easiest metric to apply, since clock time is usually sufficient. However, with the use of different parallel hardware, a given algorithm will scale differently, primarily due to the topology of the architecture, but also due to memory allocation and language attributes. Performance scaling is very important, especially in AI which has a history of solutions to "toy" problems that do not scale to real problems. In some cases, the performance question is an "order" measure, for example order O(n log(n)) algorithms are better than O(n^2). Thus, algorithms cannot be compared at a single problem or data set size, although such distinctions might not in practice matter at all, because regardless of order, some problems are limited in data or object number and even the slow algorithms might be practically fast enough. Also, we must note that some algorithms only make sense on specialized hardware, and if so, one can only look at changes in performance in light of that specialization.

III. SUMMARY

Our objective is UGV evaluation, and as such contain a system/customer requirement in addition to the science/progress requirement. Nonetheless, I have focussed on the science requirement, and from an IU perspective, rather than a "sensor" requirement, and have described three elements needed to form useful measures -- accuracy, coverage and speed. Not mentioned are the need for ground truth, testbeds, and simulators, all of which are quite necessary for useful algorithm testing.

### 4.4.7 Daniel DeMenthon, University of Maryland

Some of the goals:

- Give grades to algorithms, with the idea of just picking the best for a task.
- Make sure that an algorithm is good enough to run the vehicle safely, before it is installed on the vehicle; the output of the evaluation is just Pass or Fail.
- Understand WHY an algorithm is better than another for a task, with the idea of improving and interbreeding tested algorithms.

The three goals above probably require different types of performance evaluations. The third goal, understanding why an algorithm is better, seems to be important. The reason this goal should be considered is that many side effects can become decisive factors in performance. For example, a stereo package may look better just because of smarter image preprocessing; if we understand this,

this function could be transferred to another package which did not perform as well overall but has better feature matching. This type of analysis may require very simple synthetic images, so that only specific functions of the algorithm are exercised at one time. Ideally each of the equivalent functional modules of each package should be tested separately against each other. On the other hand the first two goals may require inputs as close as possible to the inputs in the target task. This means that the input may be a complex dynamic image sequence, and the CAUSES or failure or success will be very hard to analyze.

Levels of complexity of inputs

- Simple static synthetic images, projections of simple known 3D geometries, controllable noise.
- Real images of static scenes with known ground truth.
- Sequences of simple synthetic images.
- Sequences of video images from moving sensors in the lab.
- Tests on moving vehicle driven by data from algorithm.

Measures of success

- One approach: Measure differences between perceived world and ground truth
- However the goal is not to reconstruct the world but to navigate. The algorithm may make large errors in its estimate of the vertical position of obstacles, it does not matter if the vehicle uses only the horizontal coordinates to navigate:
- Measure differences that are significant to the vehicle safety.
- Measure differences between target task and completion of task.
- Evaluating dynamic performance should be dicussed. One could consider the history of differences between prediction and observation in a Kalman filter.

### 4.4.8 Chuck Thorpe, Carnegie Mellon University

1) LADAR: There is an extensive tech report by Hebert and Krotkov that discusses our experiences at CMU. That should be a good starting point.

2) Obstacles: Depending on the approach taken, obstacle detection may not be meaningful without a real vehicle model. We have two methods of using range data at CMU. One of them simulates vehicle traversal over the terrain, and

determines traversability at each point. For that method, obstacle detection clearly cannot be evaluated without specifying vehicle characteristics and direction. The other method is more vehicle-independent, and ends up producing a binary map of obstacle / no obstacle. That would be easier to evaluate stand-alone.

We have some experimental data from the ERIM looking at a series of cardboard boxes, ranging from 20 cm to 70 cm cubes, at ranges from a few meters to the maximum range of the ERIM. That experiment should be easy to reproduce.

We have no data on potholes, which is a critical item.

3) Overall: Ed Riseman's Vision Manifesto has lots of good ideas. E.g. in the section on road following, we have:

> Criteria for Success
>
> > A successful road following vision system will have to be fast, accurate, and long range, as specified below. All these numbers are open to change as vehicle design matures. In particular, robustness requirements will depend on other parts of the system.
> >
> > Vehicle speed. The bottom line is that we have to move fast on roads. On suitable roads, we should be able to drive the SSTV's flat out.
> >
> > Processing speed. For typical vehicles, generating 10 road models per second is faster than the vehicle control bandwidth, and is therefore fast enough. This also carries a TBD requirement for minimizing latency.
> >
> > Range. We probably need road geometry to 25 meters. (A top speed of 55 kp/h is approximately 15 m/s. Stopping the vehicle at 0.5 G deceleration, which is OK for asphalt, would require 22.5 meters to stop).
> >
> > Accuracy / smoothness. Depending on road width, sizable errors could be tolerated. But rapid changes in error will create an unstable system.

Robustness on a single road. Number of runs per error on relevant road types, mean distance traveled between errors.

Robustness for multiple roads. Number of intersections and road type switches handled per error.

Robustness for environmental conditions. Decrease in range, accuracy, robustness with unfavorable visual conditions.

### 4.4.9 Ramesh Jain, University of Michigan

In your outline of the workshop report, you have landmark recognition and surface characterization as separate topics. I think that there is a strong relationship among surface characterization, landmark recognition, and feature detection (not listed). If they are not discussed under one topic, they should be at least close in their position.

Also, I did not see any explicit mention of the images used or any methodology used to define performance measures for the characterization. I feel that you are doing an excellent job in stereo algorithms. Many issues that you considered for stereo, should be discussed in more general context in the report.

# 5. REPORT OF THE WORKING GROUP ON PLANNING

Working Group Participants:
Scott Harmon (Chair), Hughes Research Labs
Ed Durfee, University of Michigan
Carl Friedlander, ISX Corporation
Fred Garret, Martin Marietta
Dave Payton, Hughes Research Labs
Ed Riseman, University of Massachusetts
Marcel Schoppers, ADS/Booz Allen

## 5.1 Introduction

The planning evaluation working group discussions can be organized into performance evaluation structural considerations, specific metrics brainstorming and discussion of performance evaluation issues. The structural discussions examined the process of evaluating planner and plan performance. Many of the issues which evolved from these discussions were again addressed in the issues discussions. Some time was spent brainstorming to identify a list of specific performance metrics which might be meaningful.

## 5.2 Performance Evaluation Structural Considerations

The process or structure of evaluating planner performance was seen as important as the metrics which were actually evaluated. This structure determines how the data are collected, analyzed and utilized. The discussion below is not exhaustive but does represent a reasonable start in addressing the general problem of performance evaluation. Some confusion occurred during the working group interactions until the specific components of a planner were identified. This is presented below as the canonical planner components. In addition, several requirements for performance evaluation were identified. As a note, a distinction between evaluating a planner and evaluating a plan was identified as existing but the specific nature of the distinction was not resolved. Much of the following discussion (and metrics) applies both to a planner mechanism and to the product of that mechanism (i.e., the plan). However, the duality may not be universally applicable.

### 5.2.1 Canonical Planner Components

In order to facilitate unambiguous discussion of a planner and its function, the components of the planner needed to be defined. No attempt was made to design the structure or interactions between these components. Furthermore, this definition serves only the developers and any evaluation of performance which is done to serve them. The user of a planner should see it only as a black box which performs planning functions. This same definition applies to all levels of planners which might be used for UGV missions. In addition, these components are those of situated planners since they appear to be best suited to the current UGV program.

The four components which can be associated with every planner are:

> long term knowledge - i.e., operations from which the plan is constructed
>
> short term knowledge - e.g., map, world state
>
> planner algorithm - the algorithm applies the long term knowledge to the short term knowledge
>
> control heuristics - heuristics which handle such functions as conflict resolution, prioritization and pruning

Cost functions are not mention explicitly in the components above. However, they are embedded somewhere in one or more of these components. Where cost functions reside is both implementation dependent and multi-dimensional. Component definition often results in the need to define the vocabulary which defines the components. The words which are used here are not, by any means, universally accepted in the planning community. Some effort was devoted to finding the most descriptive words but that was short-circuited to maintain forward progress. In an effort to minimize any confusion due to these choices of semantics, a brief explanation or example is provided with the term.

### 5.2.2 Performance Evaluation Requirements

Evaluation of the performance of any subsystem requires metrics (i.e., the parameters which are evaluated), benchmark tests (i.e., the things that the subsystem is doing when the metrics are evaluated) and the instrumentation which

actually collects the data for the evaluation. In addition, comparison of the performance of the subsystem to some standards is often useful and can provide a means to obtain dimensionless metrics. Performance standards are often needed to identify if the subsystem can meet user-defined requirements. However, performance standards are not a necessary requirement of performance evaluation if the only purpose of comparison of two designs or successive realizations of the same design.

Metrics

The term metrics was chosen to describe the parameters which characterized a subsystem's performance to avoid the biases which are often attached to other terms. This particular choice was specific to the planning working group and not accepted by the other workshop participants. Meaningful metrics can only be identified after their purpose and ultimate audience has been defined (i.e., good metrics are user-based). Reasonable metrics enable decisions to be made regarding system effectiveness. In addition, they should make it possible to understand and define the limits of a particular subsystem approach and to identify which approaches are appropriate for which types of problems.
Metrics can be used for several purposes but should have the following general properties.

They should be unambiguously measurable.

They should be derived from the subsystem requirements.

They should be traceable to the metrics which describe the performance of the subsystem's parent system.

They should be easily understood and, thus, explainable to people other than the performance evaluator.

These are, in some cases, difficult criteria to meet (as this working group found in subsequent discussions) but are necessary for useful performance measurement.

Benchmark Tests

The term benchmark tests is adopted from computer system performance evaluation and has the same meaning. These tests consist of one or more standard

tasks which the subsystem executes during which the data on the metrics is collected. In general, these tasks should be related directly to the application of the end system. In other words, they should be achievable by the desired system. Standardized tests should guarantee that all aspects of the subsystems which are being compared are the same except those which are being tested. This requirement implies that planners must be tested with the same internal information.

The UGV tests should include both tasks which exercise the system mobility and the RSTA functions (both independently and in combination).

Instrumentation

Instrumentation provides the raw data from which the metrics can be derived (possibly, directly). Instrumentation can be placed at several places in the subsystem. It can record information which is related to the subsystem's global behavior or the behavior of any of the individual components. In describing a planner's global behavior it is treated as a black box which hides the specifics of the design and implementation. This can make the evaluation technology independent. Descriptions of global behavior can be useful to both user and developer whereas component behavior is probably only meaningful to the developer. In addition, the developer needs to evaluate a planner's performance at every decision made during task planning and execution to prevent early poor decisions from unfairly contaminating the performance during the entire task. Data on performance can be collected in a variety of ways. Parameters which describe the outward behavior of the subsystem can be collected. The internal states of the subsystem can be monitored and inspected. Finally, parts of the subsystem can be ablated and the performance of the reduced system can be observed.

Standards of Performance

Standards of performance have several purposes in performance evaluation although they are not needed for every performance evaluation circumstance. They are derived from the application requirements and can be used to ensure that the resulting systems performs adequately for the task. They can also be used to normalize the performance metrics and, thus, generate non dimensional metrics which directly reflect a subsystem's success or failure in meeting the application requirements.

For planners, one obvious choice is to compare the real planner's performance with that of an ideal or omniscient planner. This gives a clear notion of planner efficiency in a Carnot sense. Nevertheless, comparison with an optimal planner is not without its associated issues. These are discussed in greater detail below. It is also possible to compare the performance of an automated planner with that of a human executing the same task. However, this choice of standards can, in some cases, be unfair and, perhaps, not meaningful. It might be more meaningful to compare the a planner with the performance of a human executing the task as a teleoperator. In this way, the human is encumbered with the same sensing and actuation limitations as the automated planner.

## 5.3 Specific Metrics Brainstorming

Two classes of performance evaluation metrics were derived from the working group discussions. The first class occurred during the discussions and brainstorming which was unconstrained by acceptability criteria. The second class was derived subject to specific acceptability criteria and, thus, are termed constrained.

### 5.3.1 Unconstrained Brainstorming Results

During the course of the working group discussions, several planner metrics were suggested either explicitly or implicitly. No attempt was made to analyze these suggestions in order to maintain the continuity of the discussions. However, they are presented here to guarantee that every fragment of the discussion contents are captured.

No attempt has been made to define how these various parameters can be determined. No significant filtering has been done. If two metrics are similar but appeared to capture different nuances of the problem they were both included despite possible redundancy.

- coherence of multiple vehicle behaviors
- extendibility
- fault tolerance
- brittleness
- versatility
- flexibility
- effectiveness
- explainability

. sensor input requirements
. goal specification requirements
. a priori knowledge requirements
. subsystem feedback requirements
. ability to generate recognizable landmarks
. coupling between the goals and constraints
. quality of task performance
. required communications bandwidth
. required computational capabilities
. required time overhead
. sensitivity to initial state
. quality of commands generated
. difficulty to include new knowledge
. sensitivity to time limitations
. survivability with hardware lesions
. algorithm extendibility
. survivability in "strange environments or circumstances"
. rationale inspection/justification
. derivatives of competencies with respect to various variables and resources
. number of recognizable landmarks generated
. visibility of landmarks generated
. executability of plan generated
. minimized resource contention for sensors, computation, communications, effectors and other processes
. planner algorithm complexity
. plan complexity
. acceptable mission/goal complexity
. level of integration with other subsystems
. communications cost (e.g., risk)
. mean number of runs between code changes
. mean time to plan failure
. number of times human intervention is required
. resource utilization effectiveness

### 5.3.2 Constrained Brainstorming Results

Some time was spent trying to identify planner metrics which met some of the requirements which were introduced in the previous section on performance evaluation structure. Specifically, any metrics which were suggested had to be

quantitative and measurable. These constraints upon the metrics which were suggested often required that considerable discussion be devoted to the goal of reducing a qualitatively suggested metric to a well-defined measurable quantity. After some time, several categories of quantitative performance metrics evolved which could be applied along different dimensions of the planner (and plan). The ability to combine these various categories with the different dimensions of planner performance provides a combinatorial situation in which a very large number of individual metrics can be defined. No attempt was made to define these individual metrics because the effort was concentrated on discussing the general characteristics of this metric space.

Timeliness

Timeliness refers to how well the planner operates within the time constraints of the mission. Timeliness can be applied to the whole plan and to its subgoals. The specific metrics derived which were related to timeliness are:

time to generate a plan,

time required to execute,

time envelope of task, and

number of deadline failures.

Precision

The precision metric describes how accurate the plan is to accomplish the mission. It is measured by the deviation of the plan from the optimal plan results. The specific precision metrics which were suggested are:

spatial deviation from the optimal route,

errors of plan outputs in distance and time,

quality of goal achievement (e.g., how close in space and time), and

standard deviation of route errors.

Completeness

The completeness metric characterizes how much of the specified mission was accomplished. The completeness metrics which were discussed are:

number of assigned goals achieved per plan and

number of goals not achieved.

In some cases, the system may encounter situations which were not explicitly described in the mission in terms of goals. In these cases another metric of completeness is needed:

[number of situations planned]/[ number of situations encountered].

Confidence

The confidence metric represents the certainty of plan success. This is one of the most difficult of the metrics to measure but has real value in characterizing the performance of the planner to a user. This metric cannot be generated from the planner's evaluation of its own performance. It must be derived from multiple trials (i.e., experiment). This was recognized as difficult but the working group felt that much of the data which are needed to derive confidence could be collected during normal experimentation. In addition, the user will probably require characterization of the performance of the system in multiple trials anyway.

Sensitivity

This quantity represents how any other metric (the dependent variable) changes when the values of an independent parameter changes. It effectively describes variations of the planner's performance surface in the space of various parameters of interest. Several independent parameters of interest were discussed:

map resolution,
sensor resolution,
variations of all inputs to the planner,
goal specification completeness,
goal specification granularity (i.e., resolution), and

accuracies of all the above parameters.

It is possible that planner performance could also be sensitive to variations in other independent parameters but these were the only ones which were discussed in the working group.

Resource Utilization

The amount of resources which a planner consumes is also a metric of performance which is of interest, at least, to developers. This can be expressed by

[amount the resource is used (in units of time)]/[amount of the resource which is available].

Another possible metric which is related to resource utilization is the precision of resource utilization with respect to an optimal plan.

Executability of Plans

This metric expresses the reasonableness of the plans generated. It took some effort to quantify this metric. The result of the discussions about this metric was the number of unexecuted commands issued to the control systems. These are commands which could not be implemented rather than those which were not implemented because the situation for which they were intended did not occur.

Execution Coherence

Execution coherence describes how well the subsystems cooperate to achieve the global goal. The notion of coherence occurs when loosely coupled subsystems are operating independently to achieve individual goals which (theoretically) contribute to the accomplishment of one or more goals of the total system. Two sets of metrics were identified for coherence:

[performance of a distributed system]/[performance of an equivalent centralized system] and

standard deviation of coherence metric.

This metric can be applied to the individual subsystems as they contribute to a

single vehicle's performance, to the individual vehicles as they contribute to the performance of the unit and to the unit as it performs the mission.

## 5.4 Discussion of Performance Evaluation Issues

Four issues were discussed to some considerable extent: planning's scope of responsibility, planning and resource management, comparison of real plans with optimal plans, and measurement of task performance quality. These issues were identified early in the working group discussions and were addressed in greater detail later.

### 5.4.1 Scope of Planning's Responsibilities

Some concern arose in identifying the scope of the planning working group's responsibilities. This issue describes the intersection between planning and system. In general, many (if not all) aspects of system performance could be traced either directly or indirectly to planner performance. Some clear areas of overlap include:

proactive planning versus lower level control,
reactive planning versus low level control,
planning versus sensing, and
planning versus total system performance.

Other areas of overlap certainly exist. Two issues of concern were discussed:

assigning blame and

establishing a context for planning metrics.

Furthermore, confusion of scope could make comparisons between various alternatives extremely difficult. Confusion of scope of responsibility may also become important in the context of calibration.

### 5.4.2 Balance of Planning and Resource Management

This issue encompasses the resource contention problem. It becomes important in the context of limited shared resources onboard the vehicle. This issue arose when it was realized that resources are required to accomplish the planning of resources. These are the same resources which are needed for mission planning.

The balance between planning and resource management describes:

the tradeoff between plan quality and planning cost and

the execution of planning as it competes with the execution of the plan.

This balance also includes the issue of dynamic reconfiguration. Working group participants recognized that the system integrator was concerned with the larger issue of resource management and acknowledged that this is a long term problem which will require considerable resources and time to address.

### 5.4.3 Comparison of Real Plans with Optimal Plans

During discussion of the precision metric, the issue of comparing real plans with perfect plans drew the working group's attention. Real plans are those which are generated by the planning components. Plans may be generated at several layers of the whole system because it is likely that the planning components will be layered into a complete coherent planning system. Optimal plans are those which are generated by a planner with perfect and complete global information. Any reasonable planner can generate an optimal plan if it has complete and accurate knowledge of the world situation at every step in the mission (i.e., it is omniscient) even if it must resort to exhaustive search of the solution space.

Clearly comparison of a real plan with the optimal plan for the same mission can yield some useful performance information. However, when to perform such a comparison is not so clear. In fact, to avoid unfairly representing a planner's performance, the current plan must be compared with the optimal plan at the same point in the mission execution. Otherwise the real planner will be penalized for early mistakes even though the majority of the plan is good.

In addition, the influences of such parameters as

the input information (usually from the sensors),

the knowledge within the planner and planning mechanism, and

the performance of the planning mechanism

must be distinctly separable from the overall evaluation metric in order to be able

to diagnose planner problems. The working group also recognized that identification of ground truth is a difficult task. This can be measured through carefully controlled performance tests. The vehicle should be given all of the knowledge which it needs to optimally perform the mission and compared with its performance when it has only a realistic amount of knowledge.

### 5.4.4 Measurement of Task Performance Quality

At several points in the working group discussions the measurement of task performance quality was addressed. The interest is to describe in as simple terms as possible how effective the a priori plan is. However, there was also interest in examining the tradeoff between taking advantage of preferential options and being constrained by some goal criteria. One relation for task performance quality was suggested:

quality = [real plan performance]/[omniscient plan performance].

However, it is hard to compare real plans to "optimal" plans due to the dynamics of the environment and the sensitivity of a plan to the initial conditions. Furthermore, it may be meaningless to compare real plans against the best of all possible plans which has been obtained with an unrealistic degree of knowledge about the world.

## 5.5 Summary and Recommendations

During the course of the working group discussions the components of a canonical planner were identified, a list of metrics was created through brainstorming, this list was reduced to five meaningful and quantitative metrics, the open planning performance evaluation issues were closed to some extent and some recommendations were made. These results are summarized below.

Canonical Planner

The components of a canonical planner were defined:

long term knowledge,

short term knowledge,

planner algorithm,

control heuristics, and

cost functions (which are embedded somewhere in one of the other components).

These components are viewed primarily as an organizational tool rather than a means of planner specification.

Metrics Brainstorming

The working group defined many different possible metrics in unconstrained brainstorming and during the course of the discussions. However, during constrained brainstorming the metrics which were suggested were required to be well defined and measurable. The results were a smaller set of metrics, but when they are applied in combination to the planner components and to each other, the metrics evolved to a very large multi-dimensional planner evaluation space (albeit a well defined space). The working group agonized over definitions, specificity, interactions and implications and, finally, defined a set of five simpler metrics which might be more meaningful to larger group (i.e., including both users and developers).

Five Simple Metrics for Planner Performance Evaluation

The five simple metrics which were suggested were:

timeliness (both for the whole plan and its subgoals),

precision (measured as the deviation from the optimal plan),

completeness (the number of goal successes and failures),

confidence (the certainty of plan success), and

mean number of runs between code changes (which describes the learning curve of the students).

Open Issue Closure

Some form of closure (within the context of this meeting) was obtained on three issues which were raised during the working group discussions:

task performance quality - must be addressed dynamically at each step in execution;

resource contention - a tradeoff between plan quality and plan cost which is a long term problem; and

scope of the planner - confusion of scope makes comparisons difficult and blurs the assignment of blame.

Recommendations

Two concrete recommendations evolved from the planning working group discussions:

develop metrics for near term demonstrations

but, pay attention to the longer range goals of the program.

## 5.6 Individual Position Papers

### 5.6.1 Ed Durfee, University of Michigan

In this position paper, I will emphasize performance evaluation criteria for the Demo-II project with respect to the capabilities and limitations of coordinated vehicular activities. To set the stage for my statements, I first present background representing my view of the relationships between and responsibilities of the various planning components. I then look at the challenges faced in Demo-II for coordinated vehicular activity, and from these challenges derive a first cut at identifying the relevant performance parameters and benchmarks to evaluate progress at meeting these challenges. I conclude with a few remarks about the broader performance evaluation needs for planning technology.

PLANNING COMPONENTS: In the multivehicle Demo-II system, we often see a distinction between "global" and "local" planning. However, it is often unclear what aspects of planning the "global" and "local" apply to. I can think of at least 2 different aspects that seem to have been blurred: the range of VIEW and the range of CONTROL. The distinction is pretty simple: Just because a system has

a (relatively) global view does not mean that it has global control. I would characterize the planning components in Demo-II as:

- Mission planner: Has both a global view and global control. An off-line system, the mission planner plans the big picture and assigns responsibilities to the individual vehicles. However, since it is off-line, its global view is inherently limited to static models of the vehicles and their environments. Thus, robust mission plans must either predict and plan for a multitude of dynamic contingencies, or must be underspecified to give vehicles flexibility in how they carry out their portions of the mission plan.

- Coordinated vehicle planner: Part of a vehicle's local planner, this component should have a (relatively) global view but only local control. For example, a lead vehicle in a convoy should have a more global view of this fact, and of who is following it. With this view, the vehicle can intelligently direct information about obstructions and detours to only its followers; even though it cannot directly control them, it can impact their activities. More generally, the coordinated vehicle planner component is responsible for the on-line elaboration of the more abstract or conditional mission plan.

- Local vehicle planner: Also part of a vehicle's local planner, this component should have a local (to the vehicle) view and local control. Given the current objectives of the vehicle as specified by the mission planner and elaborated by the coordinated vehicle planner, the local vehicle planner further elaborates the plan into increasingly detailed steps, and is responsible for activating appropriate behaviors for accomplishing the plan. Components of this planner can include the path/route planner and the collision avoidance planner.

- Behavioral system: At another level, we could view the local planner as having global view and control of what the individual vehicle is doing. In that case, then we can view the behaviors as having local views and control

- local to small numbers of sensors and effectors.

DEMO-II CHALLENGES FOR COORDINATION: At the mission planning level, a principal challenge is to develop coordinated multivehicle plans that constrain the activities of the vehicles so that they work together well, while avoiding overly constraining the vehicles so that they become paralyzed (cannot execute the mission plan) and must await a new mission plan which must be formulated on-line. It seems that requiring mission replanning on-line should be

considered a high-cost mistake. However, having a minimal mission plan that causes substantial coordination activity on the parts of the vehicles on-line is equally undesirable. The challenge, then, is in flexibly developing mission plans that predefine a reasonable amount of the coordinated activity and that leave the vehicles with enough leeway to coordinate the rest on-line without requiring replanning at the mission level.

At the coordinated vehicle planner level, the challenge is to provide a vehicle with an appropriate model of how it could or should interact with other vehicles. If a vehicle must always represent and reason about other vehicles in great (and often unnecessary) detail, then coordination decision making will be slow and cumbersome. If the "global view" is too abstract, on the other hand, opportunities for coordination can be lost. Given global models of others at varying levels of abstraction, therefore, a vehicle can selectively elaborate on possible interactions with them and identify what "global" (non-local) information it needs (or can supply) to help it (or another) make better decisions about its current and future courses of action (such as telling followers about obstructions, or requesting a location update from a colleague in bounding overwatch, or warning another vehicle about a delay for a rendezvous planned by the mission planner). This kind of information passing is based on models of coordination; the information itself comes from sensors and from local planning.

At the local planner level, the challenge is to detail a plan that meets the objectives and constraints imposed from the levels above, and to provide feedback about actual (or near) constraint violations to forewarn the coordination component.

POTENTIAL PARAMETERS AND BENCHMARKS: Given the uncertainties of the environment, I would argue that correspondence between plans and events, or between estimated and real states, is an inappropriate measure. Instead, I would argue that what really matters is the degree to which global (mission) objectives are achieved: Have the goals been accomplished within the constraints imposed? In broad terms, the constraints might not only involve limits in time and space, but also in communication, sensing, etc., as well as dependencies on these resources. Of course, this just pushes the problem to a different level: If we say that the measure of performance is the ability to accomplish objectives within constraints, then we have to define more precisely what constitutes desirable objectives and reasonable constraints.

As for benchmarks, comparison (and competition) with human teams in contests

of skill and cunning is appropriate, with one warning. For human teams, identifying the "global" criteria for success is difficult because the criteria are inherently distributed among the players. In computational systems, we can explicitly represent the global criteria. So the danger is that the task faced by human and artificial systems might be described identically, but interpreted by those systems differently, leading to performance that might be divergent and difficult to contrast.

EVALUATION OF PLANNING: In this position paper, I've concentrated on the evaluation of coordinated vehicle performance. This is only part of the planning problem. My feeling is that criteria similar to what I have discussed are applicable at all levels of planning, however, since each level is simply an elaboration of the one above it. In fact, as noted above, the global-local distinction can be made within a vehicle as well as within a team of vehicles. However, as has been noted elsewhere, along with an elaboration with increasing localization as we work down the hierarchy, there is also reduction of temporal extent being planned/controlled. I think it will become increasingly important at more detailed levels to provide verifiable bounds on the timing of the planning components. That is, at the coordination level, deadlines are typically due to expectations between agents, and these deadlines are thus often soft and negotiable. At the behavioral control level, however, deadlines are generally more hard and based on physical rather than social constraints. Provability at that level will be much more important.

### 5.6.2 Dr. Carl Friedlander, ISX Corporation

Whether the goal is validation of operational capabilities, identification and selection of the best implementation or progress measurement as part of an experimental investigation, the evaluation of any complex software system poses a difficult but necessary part of software system management. Developing metrics with which to differentiate planners is inherently as difficult a task as developing the planners themselves. The needs for such metrics are self evident. You cannot control or improve what you cannot measure.

To make progress, one must partition the possible entities about which we wish to make decisions into at least two categories. We begin by differentiating between situated planners and those that are domain independent. We could further differentiate between skeletal planners and complete or between reactive planners and long range. In order to produce useful metrics for the Unmanned Ground Vehicle (UGV) domain, we will restrict our further comments to situated

planners only.

To be useful, metrics must satisfy several conditions. First, they must be objective in order to help reduce the effect of any external intellectual bias. As an aid to the development of objective metrics, they can be derived from either the products being measured, the process in which the products are used, or the domain in which the integrated system is applied. Metrics must be effective. To be effective, the metric must provide differentiating values that support the user in distinguishing between multiple candidates. Finally, metrics must be decomposable into or expressible as detailed domain specific measurable elements that can be addressed by instrumenting workable, supportable test cases.

A starting place for defining metrics for evaluating and comparing plans and planning components will be the identification of domain independent measures. These measures will begin with high-level specification of desirable characteristics such as workability, robustness, effectiveness, and others. The initial expansion of these top level measures will aim for measures that are still domain independent but also meaningfully detailed. In turn, these will be decomposed into specific, testable domain dependent metrics. Clearly, we must also construct metrics that allow us to measure the ability of a system to perform within a given set of limits and without exceeding specified bounds.

Metrics must also be designed to live within domain reality. For example, it does not make sense to attempt to develop a metric that measures the degree to which a vehicle avoids running over the soldier that it is escorting. On the other hand, any set of metrics that one creates for this domain must heavily reduce the acceptability score of a planner, regardless of other abilities, that does run over its soldier escort.

I believe that the only effective structure to date, with which to provide situated planner metrics, is test cases. Test cases must be further organized into the categories of behavior metrics, inspection metrics, and reduced functionality metrics. Behavior metrics are often applied indirectly and related to the observed total system behavior as it reflects the behavior of the planner. As a result, we are forced to carefully design benchmarks and data collection processes that do their best at isolating the planner from its surrounding system. Inspection metrics are based on direct measures of the output plan. Such measures as path length, number of steps, and others can be measured through direct measurement of the plan. Where known optimal solutions for direct measurements exist, comparisons with the optimal are appropriate measures. Nearly all planners can

be controlled by varying their inputs. As a result, systematic variation of input parameters and observation of the planner performance as resources become critically constrained and plan limiting make appropriate inspection metrics. Finally, reduced functionality testing involves tests where key elements of the planning system are removed or disabled. This technique is used to isolate contributions made by different planning components. Complex planning systems which mix generative strategies with algorithmic route planning and reactive navigation components must be partitioned in order to be effectively measured.

This paper argues for development of domain specific behavioral metrics that are presented as test cases, inspection metrics that are related to test cases presented as planning scenarios, and reduced functionality metrics that are again related to specific test cases for which benchmarks can be established. In addition, we argue that inspection metrics and behavioral test cases be chosen so as to extract information about the performance of the system as it approaches the boundaries of system capability.

### 5.6.3 Fred Garrett, Martin Marietta

*1) What can performance measures and evaluation for planning systems do for the ARPA UGV program?*

- Provide a means to compare planning systems
- Determine the important aspects of the planning problem from various points of view (most importantly, the end users)

*2) Which planning technologies or other aspects of planning performance need to be measured and evaluated?*

To the list of POTENTIAL PLANNING TECHNOLOGIES, add:
- task/activity planning
- payload planning

To the list of POTENTIAL PARAMETERS, add:
- reusability
- flexibility
- amount of knowledge required to execute/understand plan
- information content of plan (inverse relation to above)
- sensitivity to initial conditions
- "executability" of plan (a measure of the degree to which the plan takes into

account the physics of the mechanisms that eventually realize the plan)
- ability to cope with unknown/uncertain information

*3) For each of these, what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?*

Determine the potential role of each ARPA team member in planning performance evaluation.

POTENTIAL BENCHMARKS
Canned scenarios or live gaming environments in which both computer and human players compete against each other in complex environments in contests of skill and cunning

NO. We should focus on how human performs with/without computer assistance in higher level planners (mission, path/route, task/activity, coordinated vehicle planning). Probably do not wish to pit ourselves against human driver unless they are remote.

### 5.6.4 Dave Payton, Hughes Research Labs

After having had the benefit of reading the above papers, I would like to add a few comments of my own.

I agree with Carl that we first need to establish a set of domain independent high-level metrics. In the IU domain for example, they used metrics such as accuracy, characterization, compatibility, cost effectiveness, ease of use, effectiveness, flexibility, portability, robustness, testability, and verification. I believe that many of these metrics will apply to our domain as well.

Ed Durfee also made an important distinction between VIEW and CONTROL. We may want to construct metrics that specifically isolate tasks or functions that involve different ranges of view and control.

In evaluating planning functions for a UGV system, we should attempt to isolate tasks which we believe will require planning without making assumptions about how these planning tasks are to be performed. The outputs we measure and evaluate should be limited to those things that are available in a standard military context.

I have briefly outlined a few tests that I think would be fruitful for use in performance evaluation. These tests all involve planning in one form or another. Implicit in each test are one or more performance metrics.

User interaction with the system

- how easy is it for the user to express mission goals and requirements
- how easy is it for the user to verify that the system will operate as expected (how well does the system communicate to the user what it will do)
- how closely does what the system "plans" to do resemble what the user thought it should do
- how easy is it to alter what the system will do
- how long does it take for the user to enter a mission, plan it, and confirm its correctness
- how much communication is required

Benign environment performance

- Given no surprises, the system should do what it was expected to do. (This implies that the abstract global view is a fairly accurate representation of the more detailed local view. Any obstacles that are not in the initial mission description ar relatively insignificant to the overall execution of the mission.)
- movement (applies to each vehicle)
- make correct use of terrain
  - does it choose reasonable routes
  - does it move in a manner that is in accord with military doctrine
  - does it make effective use of cover and concealment and avoid known obstacles
  - does it enter areas that could have been avoided using a priori map data
- timing and speed
  - reach designated control measures at the required times
  - keep speed within specified limits
- observation (applies to each vehicle)
  - does it maintain observability to potential threats or targets
- coordination (applies to all vehicles as a group)
  - do they operate as a unit
  - do they maintain correct formations
  - do they minimize communication

Real environment performance

- The local environment may have obstacles and other feature that differ significantly from the abstract global view.
- All of the tests for a benign environment would apply in addition to the following:
  - movement
    - does it circumvent obstacles in a timely and efficient manner
    - does it minimize need for operator intervention
    - does it avoid unexpected areas of exposure to threats
  - observation
    - does it exploit unexpected opportunities to observe enemy
    - does it respond appropriately to new enemy sightings
  - coordination
    - do they adapt formation to disabled or delayed vehicles

This is just a first crack at some of the tests we might want to consider for evaluating performance. I would like to reemphasize my desire to avoid measures that assume particular planner implementations. We therefore must evaluate performance based only on external manifestations of the system, and not on internal representations used by the planner. Only in the case of the user interface should we be allowed to evaluate a planner on the basis of something other than actual vehicle performance.

# 6. REPORT OF THE WORKING GROUP ON RSTA

Working Group Participants:

Phil Emmerman (Chair), Army Research Labs
Tsai Hong, NIST
Martin Lahart, Night Vision Lab
Howard Stern, Robot Vision Systems Inc.
Lynne Gilfillan, Lynne Gilfillan Associates
Steve Hennessy, Martin Marietta
Jim Leonard, WL/AARA Wright Patterson AFB
John Baras, AIMS, Inc.
Richard Volpe, Jet Propulsion Lab
John Thomas, U.S. Army Material Systems Analysis Activity
Dan Dudgeon, Lincoln Laboratory, MIT
Wei Chang, Army Research Labs

## 6.1 Issues

The following issues were addressed during the workshop:

- adding acoustic sensors to UGV
- obscurity vs. occlusion
- sampling rate vs. speed of vehicle
- real time processing vs. off-line processing of images
- terrain geometry
- pos
- CC&D (camouflage...)
- operator interaction (initial conditions)
- adaptive parameters
- complexity of targets (possible inclusion of friendly targets)
- communication link
  - bandwidth and weapon for engagements
  - time constraints on firing and target verifications
- sensors:
  - LADAR, FLIR, acoustic, visible, range finder, navigation sensors, ...
- metrics:
  - confusion matrix
  - probability of identification
  - statistical latency

performance robustness
pointing
covertness
breakout of algorithms (complexity, adaptability, storage, modulability)
reliability
sensor metrics:
sensitivity

- test beds:
  simulation (synthetic imagery, correlation, validation)
  representative database (multi-sensor)
  quantify with respect to human performance
- breakdown of model algorithms (synthetic imagery)
  LAser Radar Recognition Algorithm, "LARRA" (Night Vision Lab - Wright Lab):
  need Ballistic Research Labs CAD models
  limitations: Background
  DELTAS (Nickle's Research - Wright Lab)
  Laser Plus (I Math - Wright Lab)
  FLIR:
  GT VISIT (Wright Lab)
  GTC (Wright Lab)
  PRISM (ERIM - Night Vision Lab)
- real world data bases
  NVL:
  millimeter wave (mmw) + FLIR
  LADAR + FLIR + mmw
  Wright Lab (AFWAL):
  LADAR + synthetic imagery data base
- RSTA with respect to military planners
- NAV sensors AID
- manuals on evaluation criteria for RSTA
- snap shots when platform in motion
- test beds:
  moving platform: FLIR
  stationary platform:
  FLIR w/LADAR (possible acoustic sensor)
  Alliant LADAR
  ERIM LADAR
  sensor fusion

## 6.2 RSTA Engagement Variations

Environmental:
- dust
- exhaust
- clouds
- cloud shadow
- partial obscuration and occlusion
- tree, bush, grass motion
- animal motion
- weather (fog, haze, rain, temperature, wind, humidity, etc. ...)
- background
- illumination

Target:
- signature (CC&D)
- maneuver (range, position, velocity, acceleration)
- numbers
- types (size, ...)

Robot Platform:
- maneuver (position, velocity, acceleration)
- terrain
- operator interaction

## 6.3 Evaluation Approaches

1. Develop quantitative baseline data set that sufficiently identifies status
2. Evaluate RSTA technologies:
   - stabilization, single sensors (LADAR, FLIR, acoustic, visible - NAV sensors), multi-sensor, processors, algorithms, system, communications link
3. Metrics:
   - Pd, Pf
   - accuracy of target position and velocity measurement
   - latency (statistics)
   - complexity of hardware and software (size, memory, power, etc.)
   - probability of correct classification and identification (confusion matrix, etc.)

tracking accuracy
engagement accuracy (pointing)
cost
robustness
covertness
breakout of algorithms:
complexity, adaptibility, storage, etc. ...
reliability
RSTA sensor sensitivity, resolution, etc. ...

4. Simulation:
   synthetic imagery, validation
5. Representative database (multi-sensor)
6. Quantity with respect to human performance (where applicable)
7. Ground truth

## 6.4 Individual Position Papers

### 6.4.1 Phil Emmerman, Army Research Labs

Automatic target acquisition (ATA) and recognition (ATR) are difficult problems facing the defense community. This is because there is such a wide range of environments in which ATA/ATR needs to be performed. Performing ATA/ATR in these environments ranges from trivial to virtually impossible. Some ATA/ATR algorithms work perfectly in the easy environments but not at all in the more difficult environments. Thus, ATA/ATR performance criteria must be made a function of the complexity of the environment.

A measure of scene complexity should be based on metrics such as:

- Signal-to-noise ratio,
- Target-to-background contrast,
- Percent of scene containing target-like clutter,
- Number of pixels on target, and
- Target dynamics.

These should be combined into a single measure of scene complexity. Both

automated means and human judgment can be used to determine these metrics. These should be compiled for some individual images, for use in the evaluation of algorithms that operate on only a single image, and should also be compiled for images sequences.

Simulated and real data and ground truth data should be collected for the formation of an ATA/ATR evaluation database. This data should include a wide variety of challenging situations including those which test an ATA/ATR algorithm against the following:

- Dust
- Exhaust
- Fire and smoke
- Fog, haze and rain
- Clouds and cloud shadows
- Tree, bush, and grass movement
- Target-like clutter
- Animal movement
- IR reflections
- Partial target obscuration
- Low target to background contrast
- Wide variety of target types
- Wide range of apparent target velocities and accelerations
- Wide range of apparent target sizes and aspect angles
- Large number of targets

- Moving sensor (all types of motion)

Data should be collected from at least the following sensors:

- Visible TV (monochrome and color)
- IR (variety of wavelengths)
- Acoustic
- LADAR

Simultaneous, multisensor data should also be collected, so that multisensor ATA/ATR algorithms may be evaluated.

Once the complexity of a scene has been determined, the ATA/ATR algorithm performance on that scene is evaluated. For this purpose, metrics such as the following should be used:

- Algorithm accuracy (i.e.,P(detection), P(false alarm), etc.)
- Algorithm complexity
- Complexity of hardware implementation
- Size and cost of hardware
- Hardware and software reliability
- Susceptibility to countermeasures

ATA/ATR Technical Challenges

Dust
Exhaust
Clouds & cloud shadows
Partial obscuration
Tree, bush, & grass movement
Animal movement

Low target contrast
Wide range of apparent target velocities and accelerations
Head-on motion, target stops
Wide range of apparent target sizes
Fog, haze & rain
IR reflections
Moving sensor
Large number of targets

ATA/ATR Performance Evaluation Issues

Comparison of single sensor ATA/ATR to multisensor ATA/ATR
What metrics currently exist?
What types of metrics need to be developed?
- ATA/ATR accuracy
- Algorithm complexity
- Hardware implementation complexity

How practical are these metrics?
What types of data do systems need to be evaluated?
- What sensors
- Weather conditions
- Terrain types
- Amount and types of clutter
- Obscurations
- Target types
- Target contrast
- Target velocities
- Target maneuverings
- Number of targets

How will ground truth be obtained?
Who will manage ATA/ATR database?
What medium will data be available on and in what format?
How will different systems be compared?

### 6.4.2 Martin Lahart, Night Vision and Electro-Optics

The ATR that is to be used in the UAV will use FLIR or video for region of interest detection and laser radar for target identification. Detection will make use of target motion where possible; otherwise stationary targets will be detected,

relying on target size and the contrast difference between target and background. Identification will use a model-based paradigm, in which hypotheses of specific vehicles are generated and confirmed or disproved by subsequent measurements. The algorithm will contain a verification stage, in which an operator designates a target which has been identified by automatic processes.

During the UAV demo, an end-to-end evaluation of the system will be conducted. Because the range of target and clutter conditions to be encountered by the UAV is very broad, this type of evaluation will not provide definite limits on ATR performance. The individual components of the algorithm must be evaluated under controlled conditions to determine their capabilities and failure modes. Results of component evaluations can be combined to produce an estimate of system performance, which is verified by end-to-end testing.

The state of the art in evaluation methodology must be advanced to evaluate some of the algorithm components. The area that has the greatest need for improvement is the evaluation of model-based classifiers. Some work has been done to extend the evaluation of conventional classifiers to model-based paradigms by considering the confirmation of each hypothesis as a separate classification process. The need to confirm or disprove hypotheses breaks the single classifier analysis into smaller problems that can be analyzed individually, but the methodology is similar. However, the effectiveness of the hypotheses themselves must also be assessed. In one approach that has been used, the percentage of the time each hypothesis is found to be correct is measured, and the algorithm is assumed to be slow or unreliable if it has many incorrect hypotheses. Another approach databases results of all computations and computes entropies of each decision. The assumption is that decisions with small entropies consume computational resources unwisely. The computation of entropies requires a knowledge of probabilities of decisions, and large amounts of data may be required to compute these. Both of these approaches are in early stages of development at this time.

Region of interest detections have been evaluated effectively using ATR operating curves. These are plots of detection rates against false alarm rates for different values of a decision threshold (similar to the radar ROC curve). Some work has been done to extrapolate these curves to clutter and background conditions that are different from those for which the measurements were made, but more work needs to be done to make these curves a universal tool in performance prediction. Also, the extremely small false alarm rates encountered when MTI detectors are used may make statistically reliable curves difficult to obtain. The special

advantages of this type of detector make their accurate evaluation more difficult.

The ATR in the unmanned ground vehicle is on the leading edge of the technology, as are other systems of the vehicle. New evaluation methodologies must be developed to assess it and to estimate performance.

### 6.4.3 John Thomas, U.S. Army Material Systems Analysis Activity

There are two categories of evaluation needed for TUGV; technical and operational performance. AMSAA's focus is on Technical Performance. Technical performance refers to the ability of the system to meet the technical parameters established by the combat developer in the Operational Requirements Document (ORD). Critical technical parameters for the TUGV should include the following areas:

Mobility (e.g., how fast on what terrain?)

Communications (e.g., how far/fast by what technology?)

Reconnaissance, Surveillance and Target acquisition (RSTA), Target Designation (e.g., what targets at what ranges/environments?)

Man-machine interfaces (MANPRINT) (e.g., skill level, rank, gender, etc.)

Electromagnetic environmental effects (e.g., lightning effects, static discharge, etc.)

Survivability (e.g., to what level, against what?)

Reliability & Maintainability (RAM) (e.g., mean time between failures, etc.)

Integrated Logistics Support (ILS) (e.g., levels of maintenance)

Software & Firmware (e.g., number of failures, documentation, ease of use)

Demo 2, as a part of the UGV Technology Enhancement & Exploitation (UGVTEE) program, at a minimum, will require technical performance evaluations in these critical areas. AMSAA serves as the Army's Independent

Evaluator for the TUGV. AMSAA will prepare an Independent Evaluation Plan (IEP) and a Test Design Plan (TOP) for the TUGV. The Test and Evaluation Command (TECOM) will prepare a Detailed Test Plan (DTP) which will serve as a guide for the technical test of the TUGV. As Demo 1 and Demo 2 UGVTEE products transition into a TUGV configuration, AMSAA will provide the appropriate IEP & TOP documentation to TECOM for their DTP.

As each version of TUGV (STV, DEMO 1, DEMO 2) completes a technical test, AMSAA will conduct an Independent Evaluation and document the results in an Independent Evaluation Report (IER). These findings are briefed to the Deputy Under Secretary of the Army for Operations Research (DUSAOR), other Department of the Army offices and used to assist decision makers at critical milestones.

The integration of new technologies provided by DEMO 2 requires that an analysis of operational effectiveness at the force-on-force level be made of the new system. One of the tools used to investigate the system effectiveness of the TUGV is the Army Laser Weapons Simulation (ALWSIM) model (see Fig. 1).

Another effort performed by AMSAA will be to conduct a Technical Risk Assessment of the TUGV as it progresses through its development stages. Analyses, simulations and testing will be monitored and coordinated to provide an Independent Evaluation of the risks associated with system technology, especially prior to milestone decisions.

## WHAT IS ALWSIM?

COMPUTER SIMULATION OF CLOSE COMBAT

- STOCHASTIC, EVEN-SEQUENCED
- BRIEF, INTENSE BATTLES (15-60 MIN)
- COMBINED ARMS: ARMOR, INF, A/C, AD, ARTILLERY
- UP TO BATTALION/REGIMENT SCENARIO
- DIGITIZED TERRAIN/REALISTIC ENVIRONMENT
- SMOKE & ARTILLERY DUST EFFECTS

EVALUATES BATTLEFIELD UTILITY OF WEAPON

- FUNCTIONAL MODELS: MOVEMENT, SEARCH, ACQ, ENGAGE
- ENGINEERING MODELS: ACTIVE LASER ACQ & JAMMING
- RESOLUTION TO INDIVIDUAL WEAPON SYSTEM

FIGURE 1

NIST QUESTION 1: What can performance measures and evaluations do for the ARPA UGV Program?

ANSWER: Provide answers to the following questions.

- PERFORMANCE ISSUES:
  - What are the measures of performance for an ATR System?
  - What are the measures of performance for an ATR/UGV System?
  - What is the payoff of ATR in a TA role versus recon/surv role?
  - How would the user evaluate the performance of an ATR/UGV system against specified opeerational requirements?
  - How well does an ATR/UGV function in a dirty battlefield environment?

- SYSTEMS ANALYSIS
  - What kinds of item-level analyses could be conducted?
  - What tools should be used to evaluate the overall EFF of an ATR/UGV system operating in a realistic battlefield?

NIST QUESTION 2: Which UGV technologies or aspects of UGV performance need to be measured and evaluated?

ANSWER: Demo II should provide answers to system performance for an ATR/UGV system in the mission areas of:

- Target Acquisition
- Surveillance
- Reconnaissance
- Target Designation

NIST QUESTION 3: For each of these, what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?

ANSWER: For Demo II TUGV RSTA missions, ATR technology will be used to perform the RSTA functions. Metrics and measurements will be provided for inclusion into Demo II version of the detailed test plan.

### 6.4.4 John S. Baras, AIMS, Inc.

1. Test Databases

*General*

The ATA-ATR system of the UGV is a very critical component. Many attempts in the past have led to erroneous and non-scientific bases for evaluation. We must have a family of test databases with well justified metrics agreed upon a priori, ordered according to "difficulty" as measured by these metrics. Since at the time of development of the test databases we will not have the ATA-ATR system available, the proposed metrics and their implied "difficulty" measure must be tested extensively against experienced human operators, working in the same environment as the UGV. In addition interrelationships between the proposed measures must be tested with human operators in the same way.

Using information theory and computability based complexity, we should combine the various metrics into one "compound metric" of the complexity of the test database. The implication being that the higher the value of this compound metric, the more difficult it will be to perform ATA-ATR by experienced human operators.

An appropriate sample of experienced human operators must be employed to guarantee objectivity of these subjective and perception based tests.

To limit the cost and focus these efforts on the UGV, only databases within the projected limitations of the UGV tests should be considered.

*Metrics*

Since we anticipate the use of a combination of real imagery and synthetic imagery test databases, we propose the following metrics:

Scale: number of pixels per target of interest.
Signal-to-noise ratio: target to background contrast/sensor RMS noise.
Target strength: mean target edge strength.
Number of target types.
Target image complexity: e.g., internal boundaries, uniform radiation, histogram of radiated energy or pixel values, etc.
Target occlusion: Percentage.

Degree of clutter: structure of background, statistics, etc.
Degree of target competitive clutter: e.g., clutter which looks like pieces of targets.
Image variability: is it more or less the same type of scene or wildly varying scenes?
Degree of sensor to target viewpoint variability: e.g., are all the targets flat or are some at greatly varying geometrical positions including some odd ones?
Image complexity: e.g., many targets, difficult background, occlusions, wildly varying geometries, wildly varying environmental conditions, etc.
Resolution of the acquisition system.
Range of sensor look angles.
Correlation between synthetic and real (field) databases.

The following issues must be examined carefully and justified on the basis of perceptual and human operator results. How to combine metric values for individual images over the data base? How to measure? How to represent the database statistically for meaningful statistical tests for significance, bias, analysis of variance, etc.

The above metrics are sensor specific and should be developed for each sensor used for ATA-ATR in the UGV. The same metrics must also be computed when we use several sensors in a sensor fusion set-up.

2. Test methods

Tests should be carefully designed using principles from: (a) the theory of design of experiments; (b) theory and techniques of importance sampling; (c) standard statistical tests for objectivity and analysis.

To achieve this it is recommended that the tests first be designed and tested with an experienced set of human operators. Then the tests should be applied to different groups of human operators with various degrees of competency to validate the tests themselves. Necessary adjustments should be made as needed and the process repeated until entirely satisfactory conformance is achieved.

When applied to candidate ATA-ATR systems for the UGV the tests scores should be compared to those obtained with human operators. Therefore the work just described, involving human operators, must be performed anyway.

3. ATA-ATR System Performance Tests and Measures

*General*

The ATA system must be tested and evaluated alone. The ATR system must be tested and evaluated alone. Then the combined ATA-ATR system must be tested and evaluated. These tests and evaluations must be performed for each sensor alone and for all sensor groups provided in the operational specifications of the UGV. Finally the performance of the ATA-ATR as a subsystem of the overall information system of the UGV must be tested and evaluated.

We recommend only model-based ATA-ATR approaches and algorithms.

*Metrics for the ATA subsystem*

Accuracy: how many targets detected and acquired? How many missed? How many objects or clutter detected or acquired as targets? How many out of a known set or out of an unknown set? Reaction to a new target in the scene?

Speed: Time to detect and acquire. Statistics of time as well.

Robustness: Variability of performance with respect to variations in the test database metrics.

Dependence on detailed scene information: Algorithms must not rely on a lot of scene and geometry information; like range, scale, velocity of target, target path, etc.

Dependence on off-line computations and reprogrammability: Algorithms must not depend on over-optimized lengthy computations to derive optimal on-line tests. Algorithms must be quickly reprogrammable for new target intelligence information.

Complexity: Hardware resources needed for implementation.

Graceful degradation: e.g. due to sensor loss, damage, etc.

Comparison with human operators.

*Metrics for the ATR subsystem*

Accuracy: How many targets correctly classified? How many incorrectly classified? How many missed? How many objects or clutter classified as targets? How many out of a known set or out of an unknown set? Reaction to a new target in the scene?

Speed: Time to classify. Statistics of time as well.

Economy of models used: Algorithms must use economic and efficient representations of targets.

Robustness: Variability of performance with respect to variations in the test database metrics.

Dependence on detailed scene information: Algorithms must not rely on a lot of scene and geometry information; like range, scale, velocity of target, target path, etc.

Dependence on off-line computations and reprogrammability: Algorithms must not depend on over-optimized lengthy computations to derive optimal on-line tests. Algorithms must be quickly reprogrammable for new target intelligence information.

Complexity: Hardware resources needed for implementation.

Graceful degradation: e.g., due to sensor loss, damage, etc.

Comparison with human operators.

*Metrics for the ATA-ATR system combined*

All of the above plus

Synergism: improved performance of the combined system, dovetailing design, economy in implementation.

Algorithm complexity

4. Overall Considerations

Cost
Technology selection for manufacturability and producibility
Reliability
Portability

### 6.4.5 Dr. Dan E. Dudgeon, Lincoln Laboratory, M.I.T.[1]

Reconnaissance, surveillance, and target acquisition (RSTA) will be a challenging capability for unmanned ground vehicles (UGVs). The evaluation of target detection and recognition performance is conceptually straightforward in controlled situations where ground truth can be established. The evaluation of reconnaissance performance, however, can be problematic because of the "open-ended" nature of the mission. ("Go out there and tell me what you see.") If the reconnaissance mission can be defined to be more like a detection and recognition mission, then evaluation of mission success should become easier.

The development of the RSTA capability itself has several important challenges. There must be a balance between search sensors for detection and high-resolution sensors for recognition. Performance evaluation should be considered separately for detection, which is the function of the search sensors, and recognition, which requires higher resolution. However, it is important to have the entire receiver operating curve, especially for detection. Some scenarios may require very high probability of detection; the recognition algorithms would be used to deal with the accompanying high false-detection rate.

We advocate serious consideration of acoustic sensors for target detection:
In addition to their omni-directional search capability, they are useful for detecting ground targets that are out of sight; helicopters, close air support, and other air vehicles; and artillery fire, incoming shells, and infantry fire. Acoustic sensors would be extremely useful for search to complement the imaging sensors.

"RSTA on the move" is a difficult but important technology to develop for UGVs. The difficulties lie in obtaining stabilized imagery for target detection and recognition. (Acoustics could possibly help the detection problem here, too.) Since a UGV is a dynamic vehicle in a dynamic environment, it should have the capability of processing image sequences to make a detection or recognition decision, rather than relying on a single image. Evaluation of recognition performance over many frames of imagery might involve measuring the time

[1] This work is sponsored by the Department of the Air Force

(or amount of data) needed to reach a decision, as well as the correctness of the decision.

Target detection and recognition performance evaluation methodologies have been developed by the Night Vision Lab (NVEOD), Wright Lab, and other organizations. Refinement of proven methodologies should form a baseline approach for UGV performance evaluation. The methodology itself needs to be tested and verified, however, as part of the UGV program.

# 7. REPORT OF THE WORKING GROUP ON INTEGRATED PERCEPTION/PLANNING/CONTROL SYSTEM

Working Groups Participants:
David G. Morgenthaler (Chair), Martin Marietta
Chuck Thorpe, Carnegie Mellon University
Rurik Loder, USA Ballistic Research Laboratory
Roger Schappell, Martin Marietta
Monica M. Glumm, U.S. Army Human Engineering Laboratory
Russell Watts, Cybernet Systems Corporation
James Albus, NIST
Hal Burke, U.S. Army Materiel Systems Analysis Activity
Karen Harbison-Briggs, University of Texas
Rich Luhrs, Martin Marietta
Ray Resendes, Combat Systems Test Activity
Erik Mettala, ARPA/SISTO
Jim Antonisse, MITRE Corporation

## 7.1 Introduction

This is the report of the Integrated Perception / Planning / Control Systems Working Group (IPPC-WG). IPPC-WG participants were asked to prepare prior to the workshop a position paper. Under a separate cover, IPPC-WG participants were asked to focus on the "integrated" system because other groups at the workshop would emphasize sub-elements of UGV technologies (i.e., Sensing, Planning, and ATR). The specific charter given to the IPPC-WG was:

(1) In re NIST's first question (What can performance measures and evaluation do for the ARPA UGV program?): First, although this question specifically addresses program success metrics rather than vehicle performance metrics, the integrated systems working group should address both, and that these two are not the same. Second, participants should focus their remarks on the integrated system (as implied by the first and last words of the working group's title), rather than on constituent technologies (which should be addressed by the sensing, planning, and ATR working groups). Third, participants should not limit their comments to the UGV as a mobility platform only, but consider a UGV system that encompasses all UGV-related activity within a "militarily significant" scenario (e.g., processing of mission

specifications received from a higher military command, coordinated achievement of mission goals by a multi-vehicle UGV force, ATR, post-mission debrief of reconnaissance information).

(2) In re NIST's second question (Which UGV technologies or aspects of UGV performance need to be measured and evaluated?): Again, participants should focus their remarks on the integrated system rather than on constituent technologies. Phrased another way, what aspects of integrated UGV technologies need to be measured and evaluated that will not be discovered by measurement and evaluation of the constituent technologies.

(3) In re NIST's third question (For each of these [integrated technologies, aspects, or metrics identified in NIST's question 2], what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?): The challenge here is that of finding quantitative metrics for which the measurement and evaluation process is achievable. Participants should assess the achievability of identified metric-based evaluations by identifying the key inhibitors, such as costs (ROM), logistic requirements, test setups, technical immaturity, etc.

(4) Please address also the potential role for you or your institution in UGV evaluation (NIST's workshop purpose #5).

A number of position papers were received by the chair of this working group prior to the workshop (all position papers received, prior to or during the workshop, are included in this report). The information contained in these position papers was combined by the chair to give a point of departure for working group discussions. The working group addressed NIST's questions in order; the remaining sections of this report summarize the results of the working group.

*1. What can performance measures and evaluation do for the ARPA UGV program?*

(1) What can performance measures (PMs) do for the UGV Program?

(a) PMs are a means of communication

(i) Capabilities (dimensions)

(ii) Limitations (expectation management)

(iii) Emphasis / non-emphasis

(iv) Objectivity

(v) Internal and External, at all levels

(b) PMs are a means of forecasting

(i) Evaluation of potential benefit / cost

(ii) Integration into DoD plans and requirement documents

(2) What can evaluation do for the UGV Program?

(a) Establishes a performance baseline (quantitative)

(b) Measures performance against baseline (progress)

*2. Which UGV technologies or aspects of UGV performance need to be measured and evaluated?*

UGV technologies or aspects of UGV performance fall into two categories: operational and technical. The working group developed a taxonomy of these (working from a preliminary taxonomy developed by Ray Resendes). After revising and achieving consensus on the taxonomy, the IPPC-WG (1) allocated taxons according to which working group should address them, paying specific attention to identifying those taxons that are within the purview of the integrated systems working group, and (2) discussed metrics for IPPC taxons. Table 1 gives the taxonomy and describes the metrics (or allocations) that were achieved by the IPPC-WG.

*3. For each of these, what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?*

The IPPC-WG did not have time to discuss this question, other than the notations made in Table 1.

*4. What are appropriate roles for each workshop member?*

The IPPC-WG did not have time to discuss this question. Some participants addressed this question in their postion papers.

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | Working Group Discussion / Remarks |
|---|---|---|
| Operational | Force-on-force | The IPPC-WG agreed that these aspects of UGVs were important, and were probably best measured by gaming or simulation. |
| | Improvement in loss exchange ratios? | |
| | Extended mission applications? | |
| | Reduction of personnel allocation? | |
| | Reduction of Human Exposure to Hazards? | |
| | Integrated logistics support | |
| | Reliability / Maintainability | |
| | Survivability | |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | | Working Group Discussion / Remarks |
|---|---|---|---|---|
| Technical | Physical characteristics | Physical dimensions, weight, COG | | The IPPC-WG agreed that adequate metrics exist for these aspects of UGV performance. Not discussed in detail by IPPC-WG - time limitation. |
| | | Load distribution & ground pressure | | |
| | | Automotive performance | Speed and acceleration | |
| | | | Braking | |
| | | | Steering | |
| | | | Fuel consumption | |
| | | Fording / swimming | | |
| | | Ride quality | | |
| | | Non-automotive actuator performance | Pan / tilt (mast?) rates and accelerations | |
| | | | Stability | |
| | | | Accuracy relative to vehicle coordinates | |
| | | | Accuracy relative to map coordinates | |
| | | A/C requirements | | |
| | | Electrical Power requirements | | |
| | | Vehicle Dynamics | | |
| | | Computing systems size | | |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | Working Group Discussion / Remarks |
|---|---|---|---|
| Technical (Cont.) | Command and control | Coordinated multi-vehicle command and control | Metrics should focus on things specific to one or more UGVs.<br><br>M1 = Number of UGVs<br><br>M2 = Degree of coordination, as measured by:<br>M2a = # of UGVs monitored<br>M2b = # of UGVs controlled<br>M2c = # of UGVs kept usefully busy by operator |
| | | Operator interface displays and controls | M1 = Workload, as measured by NASA's TLX<br>M2 = How long to give task<br>M3 = Error rate in giving task |
| | | Reconstitution | M1 = # people required to start-up<br>M2 = Time required to start-up<br>M3 = % time operational |
| | | Mission / route planning | Allocated to Planning WG. |
| | | World Model representational adequacy | Not discussed by IPPC-WG - time limitation. |
| | | Knowledge representation, world modeling, data fusion | Not discussed by IPPC-WG - time limitation. |
| | | Responsiveness to directed mission / route changes | M1 = Time to plan change / reconfiguration<br>M2 = Time to implement change / reconfiguration<br>M3 = Prob(success) with new configuration |
| | | Inter-operability with other military activities | Maps - Latency of collection to use |
| | | Mission debrief capabilities | Not discussed by IPPC-WG - time limitation. |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | | Working Group Discussion / Remarks |
|---|---|---|---|---|
| Technical (Cont.) | Command and control (Cont.) | Human intervention requirements | | M1 = Frequency of required teleoperation<br>M2 = Frequency of required physical intervention (i.e., at vehicle) |
| | | HFE | Operator interface displays and controls | Not discussed by IPPC-WG - time limitation. |
| | | | Compliance to MIL-STD-1472D | |
| | | | Aural detectability | |
| | | | Drive by noise levels | |
| | | | Workspace measurements | |
| | | | Demographic data | |
| | | | Time / task analysis | |
| | | | 360° ground surface visibility | |
| | Communication | LOS, quasi-LOS, and NLOS maximum distances | | Not discussed by IPPC-WG - time limitation. |
| | | Data rate, latency, frequency, bandwidth | | |
| | | Corruption / Error recovery | | |
| | | When are communications required? | | |
| | | What is the content of communications? | | |
| | | Who initiates communications? | | |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | | Working Group Discussion / Remarks |
|---|---|---|---|---|
| Technical (Cont.) | Software | Appropriate warnings and responses | | The IPPC-WG agreed to not address these aspects of UGVs at this time. |
| | | Portability | | |
| | | Documentation | | |
| | | Ease of use | | |
| | EMI | Radiated emissions | | The IPPC-WG agreed to not address these aspects of UGVs at this time because these are not currently addressed by UGV development activities. |
| | | Frequency and field intensity level susceptibility | | |
| | | Conducted susceptibility | | |
| | Environmental | Meteorological | Temperature | Not discussed by IPPC-WG - time limitation. |
| | | | Solar Radiation | |
| | | | Sand and Dust | |
| | | | Precipitation | |
| | | | Salt fog | |
| | | | Day and night | |
| | | Territorial | Rocks | |
| | | | Grass | |
| | | | Water / mud | |
| | System Performance | Degree of Automation | | Not discussed by IPPC-WG - time limitation. |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | | Working Group Discussion / Remarks |
|---|---|---|---|---|
| Technical (Cont.) | System Performance (Cont.) | Reconstitution | Setup time | Not discussed by IPPC-WG - time limitation. |
| | | | Number of setup personnel | |
| | | | Calibration / alignment | |
| | | | Tools / fixtures | |
| | | | "Constitution" (break-down time, etc.) | |
| | | Situation assessment and evaluation | Human | Not discussed by IPPC-WG - time limitation. |
| | | | Autonomous | |
| | | | Certainty / uncertainty | |
| | | | Knowledge of self-limitations | |
| | | Mobility | Path tracking | Not discussed by IPPC-WG - time limitation. |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | | Working Group Discussion / Remarks |
|---|---|---|---|---|
| Technical (Cont.) | System Performance (Cont.) | Mobility (Cont.) | Road following (and neural network performance) | M1 = Planned Path Tracking (average, standard deviation) compared with human<br>Technique: (a) drip system (b) theodolite (c) cameras on sides of vehicle<br>Uses: (1) Certainty of staying on road (2) noise model for Kalman in navigator<br>Data Sets: Video tapes are useful, but not complete<br><br>M2=Aborts per test hour (or per test mile, etc.)<br>M2a = attributable to obstacle detection failure<br>M2b = attributable to steering failure<br>M2c = etc.<br>M3 = Speed<br>M4 = Overall Mission Time<br>M5 = Smoothness (3D, including accelerations) achieved |
| | | | Off-road mobility | All of the Road-following metrics apply except M1. |
| | | | Teleoperation | Not discussed by IPPC-WG - time limitation. |
| | | | Way-point remote operation | Not discussed by IPPC-WG - time limitation. |
| | | | Obstacle recognition (objects, holes, distinguishing bushes from rocks, etc.) | Allocated to Sensing WG. |
| | | | Recognition of and response to traffic signs, signals, and moving obstacles | Allocated to Sensing WG. |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | | Working Group Discussion / Remarks |
|---|---|---|---|---|
| Technical (Cont.) | System Performance (Cont.) | Mobility (Cont.) | Coordinated vehicle mobility | M1 = Collision Avoidance Capabilities<br>M2 = Communication (1) Who (2) When (3) With good reason? (4) Did recipient receive message? (5) What (6) How much information) Avoiding Dropouts<br>M3 = Utility of information<br>M4 = Accuracy of determined relative position / attitude of friends and foes<br>M5 = Coverage provided by multi-vehicle missions<br>M6 = Graceful degradation, measured by total mission completion time included in presence of failures |
| | | | Driving efficiency, speed, dynamics, safety | Not discussed by IPPC-WG - time limitation. |
| | | Planning / Reasoning | Path Planning | Allocated to Planning WG. |
| | | | Situation assessment and evaluation | |
| | | | Risk and uncertainty analysis | |
| | | Landmark recognition | | Allocated to Sensing WG. |
| | | Surface characterization (slope, curvature, roughness, type of ground cover, etc.) | | Allocated to Sensing WG. |
| | | Degree of terrain masking (stealth) | | Not discussed by IPPC-WG - time limitation. |
| | | Mission duration | | Not discussed by IPPC-WG - time limitation. |
| | | RSTA on the move | | Allocated to ATR WG. |

*Table 1 Taxonomy of UGV Technologies and Aspects of UGV Performance, with IPPC-WG Discussion / Remarks.*

| Taxons | | | Working Group Discussion / Remarks |
|---|---|---|---|
| Technical (Cont.) | RAM-D (Reliability, availability, maintainability, durability) | | The IPPC-WG agreed to not address these aspects of UGVs at this time. |
| | Transportability | | The IPPC-WG agreed to not address these aspects of UGVs at this time. |
| | Navigation | Global absolute position accuracy | These aspects of UGV performance were not discussed by the IPPC-WG due to time limitations. |
| | | Relative position accuracy to maps | |
| | | Susceptibility to GPS dropouts or lack of landmarks | |
| | | Landmark recognition, output from | |
| | Sensors | General Characterization | Allocated to the Sensing WG. |
| | | FLIR | |
| | | Stereo | |
| | | LADAR | |
| | | Laser rangefinder / designator | |
| | | Acoustic | |
| | | Optical Targeting | |
| | RSTA | Reliability, robustness, efficiency | Allocated to the Sensing WG. |
| | | Stealthiness | |
| | | Accuracy of transmitted observation | |
| | | False alarm rate | |
| | Safety | | Not discussed by IPPC-WG - time limitation. |

## 7.2 Individual Position Papers

### 7.2.1 Rurik Loder, USA Ballistic Res. Lab

*(1) What can performance measures and evaluation do for the UGV Demo II program?*

The TWP / ARPA MOA defines the overall program objective as "to develop and mature those navigation technologies that are critical to move UGV capability from the labor intensive teleoperation to supervised autonomy." To ascertain that we have a common interpretation of this objective as a basis for our task, I am restating here the DoD definition of supervised autonomy:

> In supervised autonomy the operator retains control of only the highest level functions, which may be exercised intermittently. As a consequence, control of multiple vehicles by a single operator becomes practical, this level of autonomy has the potential for substantial force multiplication. Accompanying the increased sophistication of the machine capability is a significant reduction in communications and interface requirements. This level of autonomy enables us to partition the workload between the machine intelligence (MI) of the UGVs and the human supervisor such that the capability of both are optimally used; e.g., all routine work is done by MI. The human handles only those tasks for which the AI technologies that are required to approximate human capability to integrate and exercise control authority do not yet exist.

This interpretation, together with the implementation instructions for the MOA which contain a description of the program, provides - from a management point of view - a baseline for deliverables, hardware as well as software, and for assessing their system performance. Assessment of program success by DoD is therefore tied to the realization of their expected operational performance as an integrated system in a representative tactical environment. At program start and throughout its execution Erik emphasized and detailed this aspect of Demo II and hence I take it for granted that by now everyone in this working group is familiar with it.

In the process of establishing the program success metrics we must first, in my point of view, extract from the Demo II MOA document and from the by now sufficiently defined system a UGV system configuration and performance

baseline that satisfies DoD management expectations and allows quantitative measurement and evaluation of its functionalities with respect to this baseline. By using this approach we can establish at the same time system performance metrics.

I don't expect that we can develop sufficiently detailed Demo II performance metrics before we go into the test and evaluation phase of Demo A. However, we can set forth the system performance baseline now as well as concepts for the quantitative measurement of the individual functionalities and use the Demo A T&E results to derive reasonable performance metrics for Demo B T&E and the remainder of the program.

Availability of system performance metrics at this point would allow us to follow the program progress with respect to the DoD management expectations and, what I consider more important, provides a measure on how well we are doing with respect to the realization of the technical performance baseline we have set forth. Also, the assessment technologies and capabilities which are developed in this program will impact on how DoD will execute future robotics and hybrid robotics human development programs.

The enclosure of the NIST letter contains a listing of potential metrics and measures for our consideration. Some of them may not be applicable to the Demo II program. To ease our task, I recommend to group the listed items and others into areas that relate to certain stages of the UGV life cycle, e.g., acquisition, logistics, operation and maintenance. This allows us to set priorities in their considerations. The Demo II program is from an acquisition point of view a program that encompasses and combines elements of Science and Technology (S&T) and Advanced System Development (ASD). Since ASD requirements are more stringent, I suggest to adopt only those performance metrics and measures which relate to ASD and are also within the MOA constraints.

*(2) Which UGV technologies or aspects of UGV performance need to be measured and evaluated?*

We can generalize the benchmark tactical UGV employment scenario which Erik has put forward to accommodate a broad spectrum of possible battlefield missions and break it down into logical elements (steps, functions, . . . ). In order to discuss this and the third question in a reasonable rational manner, I am putting forward a candidate set of elements.

(1) RECONSTITUTION. Basically, this element encompasses everything that is required to prepare and keep the UGVs alert for conducting a mission. It includes bringing the UGVs into the staging area, ensuring that they are operational and ready for missions, and setting up the C3 nodal points which interconnect the UGVs with the command hierarchy and the battlefield awareness rate and supervise the UGVs.

I don't believe that we should concentrate on this element at this stage. However, it may be advisable to list its major functionalities and identify the associated metrics now in order to discuss the metrics and measures (M&M) of the other elements in the context of the M&M of the total UGV system. This would also make their consideration easier at a later point.

(2) MISSION / ROUTE PLANNING. This element encompasses all functionalities necessary to respond to a "mission directive" up to the directive to commence the UGV's mission execution. From a system point of view, we are concerned with the time, quality, complexity, and operator burden of mission / route planning for one UGV and multi-UGV missions using Demo II AI software in comparison to one human or a group of humans using fielded tools. Potential M&M are speed, efficiency, reliability, tools support, manpower reduction, cost.

Before discussing the next element, I will briefly recap the information which mission / route planning must provide to the UGVs for the supervisory oversight and control of their mission execution:

o Approximate path including bounds for each UGV;

o Time table when the UGVs are expected to be at certain predetermined locations either for conduct of sub-missions or for intermittent prescheduled communications with the C3 node(s) (e.g., for UGV position attainment verification and status report for transmitting to UGVs update of battlefield intelligence, movement of our own forces, and path / mission plan);

o Set of instructions which enable the UGVs to make decisions based on encountered situations and optimal to the achievement of the overall mission objectives; and

o Directives when and under what circumstances the UGVs must report

back to C3 node(s).

(3) MISSION / ROUTE EXECUTION. I will briefly address here the potential M&M that pertain to the navigation and mission execution by cooperating UGVs mainly from the aspect of supervisory oversight and control and with a view toward possible quantization through measurements.

(a) NAVIGATION. This is an area where I expect that most of the participants in this working group will respond. Hence, I will not elaborate on it.

(i) Reliability, robustness, efficiency

- To what degree do plans and reality correspond both in path and in time? (What is the confidence that UGVs are there where we believe they are?)

- How well can UGVs deal with surprises? (Number of unscheduled contacts of C3 node(s) due to unforeseen situations due to shortcomings in algorithms / software or due to hardware problems or limitations.)

- To what degree do software / algorithms and work partitioning and coordination with other functionalities (e.g., RSTA while driving, detection avoidance measures) impair the capability of terrain-optimal navigation and mobility?

- What is the mobility performance degradation during night time and / or under adverse weather conditions in comparison to day time navigation?

(ii) Stealthiness - detection avoidance

- Are selected paths in "critical areas" of low signature as seen from the foe's field of view? (Optimal terrain and vegetation utilization for signature suppression, communication silence breach.)

(b) COMMUNICATION

(i) Reliability, robustness

- Since the C3 node(s) and the UGVs are maneuvering systems, communication links with line-of-sight (LOS) or Quasi-LOS radios, such as the packet radios, will not be maintainable throughout the mission and, actually, are not necessary. Even though this situation is irrelevant to Demo II, we should develop M&M algorithms that allow us to assess this blackout and its consequences. Robustness of communication to countermeasures is not a part of Demo II.

(ii) Stealthiness

- Development of low electromagnetic signature hardware (e.g., direction and power adaptable antennas) is not a part of Demo II. However, we require that communication time and frequency of data transfer between the C3 node(s) and the UGVs and between the collaborating UGVs are kept minimal.

(c) RSTA. Most of the M&M aspects are probably addressed by the RSTA working group. The emphasis of Demo II work in this area is on the development of a RSTA capability for maneuvering UGVs and not for UGVs in a forward observer position (FOP). The selection of and the moving in and out of a FOP is part of RSTA on the move. RSTA on the move requires the real time meshing of RSTA and navigation functionalities. This involves coordinated processing of signals from optical and other sensors and partitioning of the processor's workload between RSTA and navigation based on the local situation. ARPA is making available to Demo II processors with the required configuration, capacity, and throughput (e.g., iWARP, IUA) through the federal high performance computing program to accomplish this very difficult task.

(i) Reliability, robustness, efficiency

- What is the consistency and trustworthiness of the extracted information and decisions that locate and identify potential threats in a highly complex and highly hostile environment? (Number of surprises, effect of adverse lighting and weather condition, effect of navigational constraints.)

- What is the extent of human involvement? (Degree of automation of the RSTA process and its capability to cope with variations in the environment and situations.)

- How efficient is the robotic RSTA process in comparison to a manned one in a variety of threat environments from benign to very hostile ones?

(ii) Stealthiness

- What is the probability for the UGVs to be detected by hostile forces compared to manned vehicles? (Signature due to active sensors, exposure due to employed tactics.)

(d) SUPERVISORY OVERSIGHT. One objective of the DoD Robotics Program is to identify, develop, and prepare for development applications those AI technologies that will substantially reduce the "manual" human involvement in and enhance the conduct of routine operational functions. Our current technology base does not yet support the development of highly robotic UGVs for tactical battlefield deployment. Hence, there will still be a substantial human involvement in their "autonomous" operation. We have to develop a M&M concept that allows us to assess the amount, complexity, and intensity of the human interaction in comparison to a non-robotic baseline system (e.g., a purely teleoperated system such as the STV or a teleassisted system such as the Demo I HMMWVs) and to follow their progression toward a true autonomy. We can break down a tactical UGV operation into potential human interfaces and identify M&M parameters that relate to what the human does during the interaction and how long the job takes.

(e) Degree of automation (time, amount, complexity, intensity of human involvement compared to baseline).

(f) Manpower reduction of robotic UGV system in comparison to manned and teleoperated baseline systems (we can only address mission / route planning and execution).

*(3) For each of these, what are appropriate metrics, performance measures,*

*measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?*

(1) Supervisory Oversight - What parameters can we measure during mission / route execution?

(a) Planned Interfaces for UGV(n), n=1, 2, . . . , N

(i) Scheduled by mission / route plan

o number of actual contacts

o time of interaction

o location of requestor and addressee

o type

o intensity of human involvement (measurement of body functions, e.g., sweat, brain activity)

o waiting time of UGV for mission continuation

(ii) Due to the occurrence of reportable events

. . .

(b) Unscheduled Interfaces initiated by Operator

(i) Change of mission

. . .

(ii) Lack of operator discipline

. . .

(c) Unscheduled Interfaces initiated by UGV(n), n = 1, 2, . . . , N

(i) Requesting Assistance for navigation

Number of contacts

. . .

Number of contacts resulting in mission adjustments

(ii) Requesting assistance for RSTA on the move

. . .

(iii) Requesting assistance for "mission"

. . .

(iv) Due to other problems

Apart from the measurement of "intensity," the above parameters can be recorded as an intrinsic part of the operator interface software, with appropriate adjustments in the communication information content. The data evaluation is straightforward.

(2) Operational Performance - What parameters are readily accessible to measurements during mission / route execution but are not included in above?

(a) Interfaces attempted by the UGVs but not realized because communication link to C3 node(s) or to neighbor UGV was interrupted

(i) Class (e.g., unscheduled . . . )

Requestor

Addressee

Time

Location

Reason

Action carried out thereafter to continue mission

(b) Interactions between individual UGVs

Requestor

Addressee

Time

Location

Reason

Action carried out thereafter to continue mission

This information can be obtained through software implementation on the UGVs and the data are directly accessible to automated statistical analysis. Results properly packaged can be used to assess the progress towards automation. However, they do not provide a measure of how well we are doing with respect to our current tactical capability. We need a basis for comparison and this basis does not exist yet.

The planned Demo II OT&E at Ft. Hood could be conducted such that it will yield a comparison of the Demo II UGV system with a representative manned system which carries out the same mission in the same environment and under the same conditions as the UGV system. This OT&E exercise has to measure parameters which are common. These should include:

(i) man-hours, active and passive

(ii) hardware used

(iii) operation time(s)

(iv) communication time(s)

(v) accuracy of transmitted observation (e.g., detection of reportable objects, their location and movement, their identification as foe or friend)

(vi) stealthiness at critical phases and areas as observed from potential foe positions (integration of observed area over time)

(vii) timeliness of execution with respect to plan

(viii) responsiveness to directed mission / route plan changes

The measurement system must be set up to allow the partition of the OT&E exercise into logical operational segments and their evaluation as subsystems. The measurement / monitoring of the observables can be done with currently available instrumentation and recording equipment and be automated. However, we would have to develop the logic and software for the evaluation of the collected information.

(c) Technical Performance. I am addressing only those T&E aspects which pertain to our working group and are conducted outdoors.

(i) Road Driving - There is practically no difference between civilian and military deployment, except that military vehicles may have to drive without "light" during low visibility. We need a very high confidence level that the UGVs can autonomously and safely drive on multiple and single-lane roads together with manned vehicles and minimal route instructions. We are working toward this goal. For Demo II we can relax the human safety aspect and limit the autonomous driving of the UGVs to "traffic controlled" road segments where we can check out the correctness of individual recognition and decision algorithms and their robustness to environmental changes.

o road following

o recognition of traffic signs / signals and responses to them

o recognition of road crossings and responses to them

o acquisition of obstacles and their avoidance

o encounter of traffic and responses to them

Apart from human observation, we can use kinematic GPS and optical sensors on the vehicle and optical surveillance instrumentation strategically stationed along route segments to monitor the vehicle performance over prolonged test periods.

(ii) Path Tracking - We can use the same "traffic controlled" road segments and measurement setup to check out the accuracy of path following routines in retracing and retrotraversing a path previously driven with and without use of the road driving software and test their robustness to environmental changes.

(iii) Off-road Mobility - Depending on the prevailing operational mission constraints, we may have different technical performance needs for off-road mobility and, thus, require special tailoring of instrumentation / measurements.

o Navigating a preplanned "fuzzy" route segment from location A to location B without RSTA and stealth requirement. Here, the objective of the UGV is to find an optimally drivable path within the route corridor selected by the mission / route planner as a passageway. The T&E has two parts: (1) Assessment of the selected corridor with respect to other passageway alternatives, and (2) assessment of the driven path with respect to other alternatives within the selected route corridor. We can vary the corridor and its boundaries and record vehicle travel time, driven distance, and acceleration history. From the latter we can determine the number of shocks exceeding a reference amplitude level and compute their energy content integrated over the vehicle path. This allows us to evaluate our algorithms quantitatively with respect to path optimization and wear and tear. We can use instrumentation which is already on-board the Demo II UGVs.

o Navigating a preplanned "fuzzy" route segment from location A to location B without RSTA but with stealth requirement. We have to superimpose the concealment conditions and measurements on the above. In the case of concealment of the UGV movement from

hostile ground forces operating in a predefined area, we can set up optical, electromagnetic, and acoustic sensors at strategically located enemy positions and search for and monitor the UGV signature with respect to background noise. We have to derive a measure which enables us to quantitatively correlate path optimization, exposure minimization, and mission / route constraints.

o Navigating a preplanned "fuzzy" route segment from location A to location B with RSTA but without stealth requirement. We may have to consider a multitude of subscenarios in the development of appropriate algorithms. We can start with the simplest case - preplanned locations for in-position RSTA along the passageway corridor for move-stop-RSTA-move, and gradually introduce true RSTA on-the-move. In my judgement, the required instrumentation and measurement techniques are here but the test setup may be a little tedious and the measurements will be time consuming.

o Navigating a preplanned "fuzzy" route segment from location A to location B with RSTA and stealth requirement. This is the most demanding task for the algorithm / software developers. It requires constant coordination and work partitioning between "stealthy" navigation and RSTA functionalities of the individual UGV as well as of all the UGVs collaborating in the mission. Also, from a measurement point of view the complexity to simulate the hostile environment is increased considerably. We have to simulate hostile forces, resting and on the move, in order to check out the algorithms which control the interface of RSTA and navigation and the vehicle response. Again, the required instrumentation and measurement techniques are here; but the test setup is complex and the measurements will be time consuming.

### 7.2.2 Jim Albus, NIST

*(1) What can performance measures and evaluation do for the UGV Demo II program?*

UGV program success is dependent on UGV vehicle and system performance metrics.

Performance measures and evaluation can provide quantitative methods for

estimating the capabilities and limitations of unmanned ground vehicles under a variety of circumstances. This is crucial to predicting the effectiveness of UGVs under future battlefield conditions, and thus for evaluating the potential benefit of UGVs to the armed forces of the nation. Military planners must understand precisely what the performance characteristics of UGVs really are if they are to integrate UGVs into the plans and requirements documents that are needed for instituting weapons systems procurements. Without DoD requirements documents and procurement plans, UGV research budgets will remain relatively small and unpredictable.

*(2) Which UGV technologies or aspects of UGV performance need to be measured and evaluated?*

The aspects of integrated UGV system technology that should be addressed include, but are not necessarily limited to:

- o path tracking
- o road following
- o off-road mobility
- o obstacle avoidance
- o coordinated multi-vehicle command and control
- o operator interface displays and controls
- o communications systems
- o navigation and path planning
- o object / landmark recognition and tracking
- o road / terrain surface characterization
- o risk and uncertainty analysis
- o world model representational adequacy

*(3) For each of these, what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?*

*(3a) For each of these, what are appropriate metrics and performance measures?*

Appropriate performance parameters include, but are not necessarily limited to:

- o efficiency
- o speed
- o reliability or robustness
- o accuracy or precision
- o degree of success in goal achievement
- o probability of failure
- o risk, cost, benefit
- o correlation between plans and events
- o correlation between perception and reality
- o surprises (differences between plans and events, or between perception and reality

*(3b) For each of these, what are appropriate measurement techniques?*

Appropriate measurement techniques include:

- o gaming scores
- o race results
- o test course measurements
- o benchmark test measurements

Measurements techniques need to be developed for both simulation and real testing environments.

*(3c) For each of these, what are appropriate evaluation criteria, benchmarks, test data, and testbeds?*

Appropriate test beds include:

- o gaming simulation test beds such as SIMNET which allow numbers of UGVs to be introduced into a simulated battlefield which includes both manned and unmanned vehicles and systems.
- o calibrated test areas such as Churchville which allow UGVs to be operated over known terrain where their movements can be precisely measured.
- o test tracks where speed and stability can be evaluated under controlled stressful condtions.
- o test data combined with ground truth so that algorithms and subsystems can be evaluated under precisely controlled laboratory conditions.

*(4) Potential role for NIST*

- o Facilitate consensus through workshops and communications
- o Help define performance parameters, metrics, measurement techniques, evaluation criteria, and benchmarks
- o Serve as a clearing house for test data
- o Contribute test data
- o Contribute test beds

### 7.2.3 Ray Resendes, Combat Systems Test Activity

*(1) What can performance measures and evaluation do for the ARPA UGV program?*

Test and evaluation assures:

(a) You don't waste money. It is the means to quantitatively and objectively evaluate your system and its components. Metrics are the language used to express the system performance. The results of T&E are the justification to determine if your system fills a military requirement.

(b) You don't hurt anybody. Autonomous and semi-autonomous machinery functioning in the vicinity of personnel is a significant safety hazard. Early and continuous test and evaluation permits hazards to be identified and remedied early in hardware development cycle when the impact on cost and schedule is minimal. It also eliminates the need for burdensome procedures that may need to be implemented during the demos to remedy hazards that cannot be fixed with hardware. This will permit a much more professional and impressive demo.

(c) You don't re-invent the wheel. In a robust implementation testing and good record keeping will prevent you from following others down a path that leads to failure.

(d) Your demo is not just smoke and mirrors. The data and analytical assessment of the system provides a technical backbone for your demo accomplishments. This is a baseline of your system's performance and provides an avenue to compare the technology with that of past and future endeavors.

(e) Tech Transfer occurs. The players in the robotics community come and go. Technical testing gives the technology maximum exposure in the government community. This assures that what we have paid for gets put to maximum usage.

Testing ultimately provides for necessary design data and provides data to answer the decision makers' basic concerns of: Will the system perform as it is supposed to? Can the soldier use it? Can we afford it?

Testing represents a small fraction of the total project cost and is cheap insurance when compared with discovering materiel shortfalls on demo day. Technical

testing provides a valuable service by helping to find and fix problems before they become show stoppers.

*(2) Which technologies or aspects of UGV performance need to be measured and evaluated?*

*(3) For each of these, what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?*

At the system level the UGV is normally broken up into three major components:

(1) Physical Characteristics

(a) Basic measurements and weight

(b) Center of Gravity

(c) Load distribution and ground pressure

(2) Command and Control

(a) LOS Maximum distance

(b) NLOS maximum distance

(c) Function check and response time

(3) Data Link

(4) Software Assessment

(a) Appropriate warnings and responses

(5) EMI

(a) Radiated emissions

(b) Frequency and field intensity level susceptibility

(c) Conducted susceptibility

(6) HFE

(a) Compliance to MIL-STD-1472D

(b) Aural detectability

(c) Drive by noise levels

(d) Workspace measurements

(e) Demographic data

(f) Time / task analysis

(g) 360 ground surface visibility

(7) Environmental

(a) Hot

(b) Cold

(c) Solar Radiation

(d) Sand and Dust

(e) Humidity

(f) Rain

(g) Snow and Ice

(h) Salt fog

(8) System Performance (Mission dependent)

(a) Day and night trials

(b) Urban

(i) inside

(ii) outside

(c) Off-road

(i) field

(ii) woods

(d) On-road

(i) Primary paved

(ii) Secondary

- improved gravel
- Belgian block

(iii) Cross-Country

- Level
- Hilly

(9) RAM-D

(a) Collect RAM data during all other testing

(10) ILS

(a) Collect maintenance data during other testing

(11) Transportability

(12) Automotive Performance

(a) Speed and acceleration

(b) Braking

(c) Steering

(d) Fuel Consumption

(e) Standard Obstacles

(f) Gradeability and Sideslopes

(g) Fording

(h) Swimming

(i) Ride Quality

(13) Actuator Performance

(14) Navigation

(a) Reaction

(b) Self position error

(c) Horizontal position error

(d) Way-point azimuth

(15) Sensors

(a) General

(i) Sensor Characterization

(ii) MMI

(b) FLIR

(i) Minimum resolvable temperature

(ii) Spatial frequency vs average temperature

(c) Laser Rangefinder / Designator

(i) Boresight Error

(ii) Moving Targeting Accuracy

(iii) Static Targeting Accuracy

(d) Acoustic

(i) Average detection distance

(ii) Noise source localization

(e) Optical Targeting

(i) Resolution

(ii) Acuity

### 7.2.4 Hal Burke, U.S. Army Materiel Systems Analysis Activity

There are two categories of evaluation needed for TUGV: technical and operational performance. AMSAA's focus in on Technical Performance. Technical performance refers to the ability of the system to meet the technical parameters established by the combat developer in the Operational Requirements Document (ORD). Critical technical parameters for the TUGV should include the following areas:

(1) Mobility (e.g., how fast on what terrain?)

(2) Communications (e.g., how far / fast by what technology?)

(3) Reconnaissance, Surveillance, and Target acquisition (RSTA), Target Designation (e.g., what targets at what ranges / environments?)

(4) Man-machine interfaces (MANPRINT) (e.g., skill level, rank, gender, etc.)

(5) Electromagnetic environmental effects (e.g., lightning effects, static discharge, etc.)

(6) Survivability(e.g., to what level, against what?)

(7) Reliability & Maintainability (RAM) (e.g., mean time between failures, etc.)

(8) Integrated Logistics Support (ILS) (e.g., levels of maintenance)

(9) Software & Firmware (e.g., number of failures, documentation, ease of use)

Demo II, as part of the UGV Technology Enhancement & Exploitation (UGVTEE) program, at a minimum, will require technical performance evaluation in these critical areas. AMSAA serves as the Army's Independent Evaluator for the TUGV. AMSAA will prepare an Independent Evaluation Plan (IEP) and a Test Design Plan (TDP) for the TUGV. The Test and Evaluation command (TECOM) will prepare a Detailed Test Plan (DTP) which will serve as a guide for the technical test of the TUGV. As Demo I and Demo II UGVTEE products transition into a TUGV configuration, AMSAA will provide the appropriate IEP & TDP documentation to TECOM for their DTP.

As each version of TUGV (STV, Demo I, Demo II) completes a technical test, AMSAA will conduct an Independent Evaluation and document the results in an Independent Evaluation Report (IER). These findings are briefed to the Deputy Under Secretary of the Army for Operations Research (DUSAOR), other Department of the Army offices and used to assist decision makers at critical milestones.

The integration of new technologies provided by Demo II requires that an analysis of operational effectiveness at the force-on-force level be made of the new system. One of the tools used to investigate the system effectiveness of the TUGV is the Army Laser Weapons Simulation (ALWSIM) model (see Figure 1).

Another effort performed by AMSAA will be to conduct a Technical Risk Assessment of the TUGV as it progresses through its development stages.

Analysis, simulations and testing will be monitored and coordinated with system technology, especially prior to milestone decisions.

## WHAT IS ALWSIM?

COMPUTER SIMULATION OF CLOSE COMBAT

- STOCHASTIC, EVEN-SEQUENCED
- BRIEF, INTENSE BATTLES (15-60 MIN)
- COMBINED ARMS: ARMOR, INF, A/C, AD, ARTILLERY
- UP TO BATTALION/REGIMENT SCENARIO
- DIGITIZED TERRAIN/REALISTIC ENVIRONMENT
- SMOKE & ARTILLERY DUST EFFECTS

EVALUATES BATTLEFIELD UTILITY OF WEAPON

- FUNCTIONAL MODELS: MOVEMENT, SEARCH, ACQ, ENGAGE
- ENGINEERING MODELS: ACTIVE LASER ACQ & JAMMING
- RESOLUTION TO INDIVIDUAL WEAPON SYSTEM

FIGURE 1

NIST QUESTION 1: WHAT CAN PERFORMANCE MEASURES AND EVALUATION DO FOR THE ARPA UGV PROGRAM?

ANSWER: PROVIDE ANSWERS TO THE FOLLOWING:

A. TECHNICAL PERFORMANCE ISSUES:

HOW WELL DOES THE DEMO II VERSION OF THE TUGV MEET ITS TECHNICAL REQUIREMENTS FOR THE FOLLOWING CRITICAL TECHNOLOGY AREAS:

- COMMUNICATIONS
- MOBILITY
- RSTA
- MANPRINT
- SOFTWARE / FIRMWARE
- ELECTROMAGNETIC ENVIRONMENTAL EFFECTS
- INTEGRATED LOGISTICS SUPPORT
- RELIABILITY / MAINTAINABILITY
- SURVIVABILITY

B. OPERATIONAL EFFECTIVENESS ISSUES:

HOW WELL DOES THE DEMO II VERSION OF THE TUGV MEET THE OPERATIONAL REQUIREMENTS STATED IN THE OPERATIONAL REQUIREMENTS DOCUMENT FOR:

- IMPROVEMENT OF LOSS EXCHANGE RATIOS
- EXTENDED MISSION APPLICATIONS

- REDUCTION OF PERSONNEL ALLOCATION

- REDUCTION OF HUMAN EXPOSURE TO HAZARDOUS MISSIONS

NIST QUESTION 2: WHICH UGV TECHNOLOGIES OR ASPECTS OF UGV PERFORMANCE NEED TO BE MEASURED AND EVALUATED?

ANSWER:

o EVALUATION EMPHASIS SHOULD BE PLACED ON THE TECHNOLOGY APPLICATIONS THAT DISTINGUISH DEMO II AS A SUPERVISED AUTONOMOUS SYSTEM AS CONTRASTED TO A TELEASSISTED SYSTEM (DEMO I) OR A SURROGATE TELEOPERATED VEHICLE (STV) SYSTEM.

o CRITICAL TECHNOLOGY APPLICATION FOR DEMO II ARE:

- MISSION PLANNING

- PATH PLANNING

- NAVIGATION

- OBSTACLE AVOIDANCE

- RSTA

o THE FOLLOWING FUNCTIONS PERFORMED BY THE SUBSYSTEMS THAT ACCOMPLISH THESE TASKS REQUIRE STATE-OF-ART TECHNOLOGIES, I.E.,

- ARTIFICIAL INTELLIGENCE

- MACHINE VISION

- MACHINE COGNITION

- NEURAL NETWORK COMPUTING

- STEREO VISION

o EACH SUBSYSTEM THAT UTILIZES THESE TECHNOLOGY BASE PRODUCTS TO PROVIDE A SUPERVISED AUTONOMOUS FUNCTION IS A CRITICAL PART OF THE DEMO II SYSTEM AND MUST BE EVALUATED THOROUGHLY AT THE "OPEN LOOP" LEVEL AS DEFINED BY R. BROOKS

o ONCE ALL SUBSYSTEM "OPEN LOOP" LEVEL PERFORMANCE SPECIFICATIONS HAVE BEEN MET, "CLOSED LOOP" FULL-UP SYSTEM LEVEL TESTS NEED TO BE CONDUCTED TO INSURE THAT UNDESIRABLE INTERACTIONS BETWEEN SUBSYSTEM FUNCTIONS ARE NOT PRESENT.

NIST QUESTION 3: FOR EACH UGV TECHNOLOGY OR ASPECT OF UGV PERFORMANCE WHAT ARE APPROPRIATE METRICS, PERFORMANCE MEASURES, MEASUREMENT TECHNIQUES, EVALUATION CRITERIA, BENCH MARKS, TEST DATA AND TESTBEDS?

ANSWER:

o FOR THE CRITICAL TECHNOLOGY AREAS DEFINED IN SECTION A OF QUESTION (1) THE AMSAA TDP AND THE TECOM DTP ADDRESS THESE TOPICS FOR THE SURROGATE TELEOPERATED VEHICLE (A SIMILAR APPROACH WILL BE REQUIRED FOR THE DEMO II VARIANT OF TUGV)

o FOR A QUANTIFICATION OF MEASURES OF OPERATIONAL EFFECTIVENESS DEFINED IN SECTION B OF QUESTION (1), THE UTILIZATION OF ALWSIM COULD BE USED.

### 7.2.5 Jim Antonisse, MITRE Corporation

*(1) What can performance measures and evaluation do for the ARPA UGV Program?*

Performance measures and evaluations mark the progress of a project internally and in comparison with related projects elsewhere, help focus attention on areas of potential difficulty and opportunity, and serve to publicize results. The former two areas are the usual targets of measures of performance and evaluation. They will receive some attention here. However, this section empha-sizes the latter role.

There are several communities whose perceptions are critical to UGV's long-term success. These are the technical community, DoD, and Congress (with the latter influenced, sometimes strongly, by the public's perceptions). Each has expectations of robotics in general and of UGV in particular. Each perceives the program in its own terms through its own requirements. Measures adopted should facilitate the assessment of UGV from the vantage of each of constituency.

Technical Community: Measures must delineate the technical aspects of integrating the UGV component technologies. They must also clearly state the issues of integrating divergent techniques within the component areas. Areas of overlapping functionality need particular attention so that trade-off studies can be made along the dimensions of, e.g., modularity vs. integration or unification of functionality in a single module. Although measures of most interest in the technical community are those that apply across tasks, in UGV measures and evaluations must be derived from military missions and be task-relative.

DoD: Evaluation of robotics technology for military missions will have to consist of mission- (i.e., task-) relative measures of performance (MOPs). MOPs should be directly related to technical trade-offs and include, e.g., trade-offs among power consumption, processing strategies, and functionality. Moreover, the DoD constituency has critical concerns that aren't apparent at the technical level. These include the integration of new systems with existing organizations and technology infrastructures, with standard operating procedures, and with training procedures for personnel who will use and maintain the system. These realms of DoD concerns will bear crucially on the ultimate success of UGV as a military system.

Congress: Measures relevant to Congressional funding committee members are likely to be derivative of DoD concerns. However, a new consideration at this

level is the potential of UGV technology for non-DoD application. The public will have an indirect, though possibly dramatic, impact in this area in its interest in non-military payoffs for robotics and in pressure for force reductions in the military.

The measures and benchmarks adopted by UGV will have a critical role in providing insight to the various interested communities of the structure and progress of UGV. The criteria adopted will largely determine how UGV is viewed and with what degree of success it is perceived. For the technical communities these criteria can lead to proper appreciation and adoption of the UGV research community's ideas, methods, and products. To that end, the performance and evaluation metrics themselves should be propagated to other research community groups, such as the NASA Intelligent Agents Benchmarks group. For the non-technical communities, the most important aspect of the performance and evaluation metrics is likely to be their role in anchoring expectations of the UGV program and related robotics work. Hopefully these will partially replace the benchmarks set by Hollywood.

*(2) Which UGV technologies or aspects of UGV performance need to be measured and evaluated?*

Performance will be assessed by communities with differing concerns, so metrics need be designed to effectively address a spectrum of trade-off issues. In the UGV program, the goal is to provide the US Army/Marines with increased tactical capability using robotic vehicles. Therefore, metrics of performance must start from the requirements of this user. Technology-oriented measures need, in principle, be derivable from metrics that matter at the military user level, and political/funding-oriented measures abstracted from that level.

The technologies to be integrated are sensors, planning, and control. "Sensors" include the software elements of automated perception. "Planning" is here viewed as the task-level, i.e., the linguistic, specification of the system's behavior. "Control" is the mapping of intended system behavior to actuator settings (acceleration, wheel turn, head tilt values, etc.). In the control arena, the intended system behavior may be expressed at the task level, as in universal plans, or as a set of state equations as in control theory approaches.

A pairwise analysis of integration issues among the technologies is appropriate (the assumption is that integration is associative). Below is a table illustrating such an approach. Integration among the technologies within one vehicle may not need

to be fully worked out before higher-order integration occurs. (This follows if an approach to system development is taken similar to Pilot's Associate, where apprenticeship is part of the development path.) Higher-order integration is among the user and the vehicle and among numerous vehicles. One such table should be constructed for each consumer of performance and evaluation results. The table for DoD should be constructed first. Once an entry in the table is crystallized its implications for other entries may be propagated. Work can begin, for instance, in deriving engineering-level measures as soon as a DoD entry is set.

An example of the way in which this table might be filled out follows. Consider sensor/sensor integration. The critical requirement is to measure the costs and benefits of integrating these systems. Suppose one sensor system is tied directly to the navigation system, and another to the land-mark system. Then the cost of unifying the systems includes the reduced dwell time on either task, the development cost of the control algorithms for deploying one sensor for multiple tasks, and the computational cost of running the controller itself. The benefits include the elimination of a sensor and the resultant reductions in power, weight, and size of the sensor package. The ability to measure these, indeed even the clear articulation of the trade-offs, could lead to important shifts of emphasis in the ongoing design of UGV. An outcome in this case could, for instance, be the serious active perception with traded control between local and global navigation.

*(3) For each element identified in question 2, what are appropriate metrics, performance measures, measurement techniques, evaluation criteria, benchmarks, test data, and testbeds?*

A full answer to this question is beyond our resources to answer. A good start will hopefully emerge from the NIST workshop. What has been attempted above is the development of a framework to organize the results of those discussions. However, two points should be raised with respect to this question. These deal with the technical work needed to develop a good set of integration metrics, and the source of the criteria themselves.

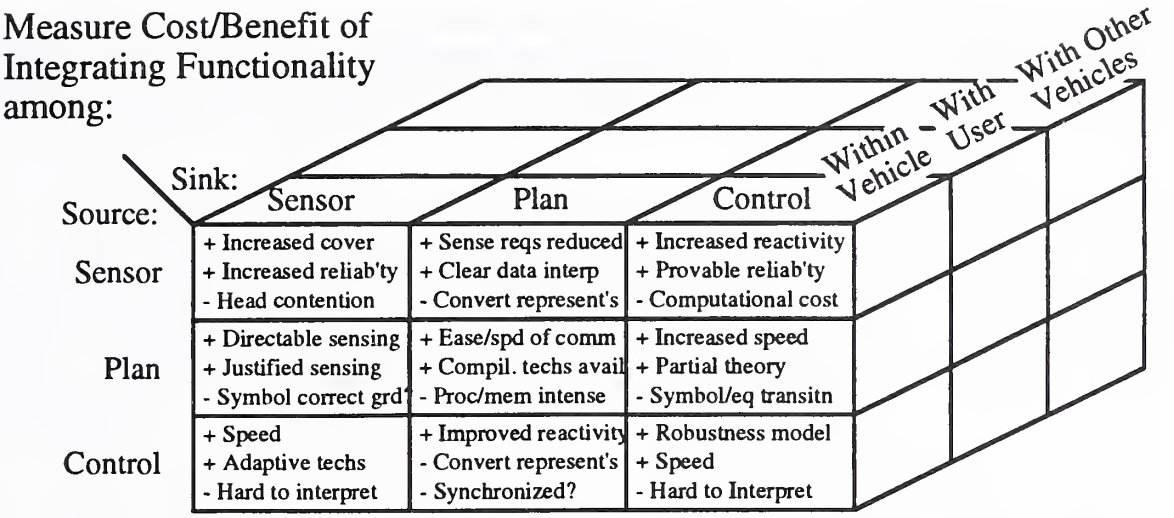
Potential areas of overlap among the component technologies must be clearly articulated. This is a prerequisite to achieving useful measures of integration. The domains of planning and control seem co-extensive, so it should be possible to develop reasonable trade-off measures in this area. The close relation of sensing to planning and control is less fully appreciated. The major trade-off among them is between sensing to achieve information about world state and computation over a world model. For instance, the more a modelling approach can be made to

work in task domain, the greater the leeway in trading the computation of present state and its sensing against one another.

The table developed in (2) is only a nominal indication of integration measures. The entries should be describable in terms of the standard "-ilities" used to evaluate system utility. A recent study of C3I systems included Interoperability, Securability, Producability, Maintainability, Reliability, Reusability, Flexibility, and Mobility. For long run, development criteria of operational UGV's might best be found by analyzing the field manuals (e.g., FM 17-15) describing the standard procedures for humans in the missions for which these machines are being developed, and planning for piecemeal automation of existing functionality.

*(4) What is the potential role for MITRE in the UGV evaluation?*

MITRE's anticipated role in UGV is in the technical aspects of UGV transition to the military. In the programmatics arena this includes the acquisition and integration of emerging autonomous vehicle capabilities for military use. This means, among other things, integration of robotic machinery into an existing technical and organizational military infrastructure. In the technology arena, this role includes work in (1) the application of planning techniques to field robotics, (2) the integration of real-time computer vision and task-based planning, and (3), the coordination of multiple planning agents in adversarial environments.




### 7.2.6 Russell Watts, Cybernet Systems Corporation

Let us make the assertion that the objective of IPPC in the case of Demo II is to facilitate force multiplication by the use of moderately intelligent semi-autonomous vehicles. Then a measure to assess the performance of the IPPC must quantify the force multiplication achieved. If we were to look at this as a computational problem then there would be two extremes with a continuum in between, with the two extremes being a totally sequential process and a totally parallel process. In computation, the conceptual equivalent of force multiplication is scalability. If we add a second processor do we accomplish twice as much work - two processors, three times as much work, etc.? The unfortunate answer to this question is that it depends on the algorithm. Therefore, if we hope to plumb the limits of this system, we must be careful to implement an algorithm that allows for a certain amount of parallelism, otherwise part of the benefit of having separate semi-autonomous vehicles will be lost.

Continuing this line of thought, we also have a certain amount of overhead in the form of an OCU and a human operator. Since the operator provides a supervisory / tasking role, it should be apparent that the less the operator must do in an operational sense, the greater the scalability or force multiplication that will be possible. This principle should manifest itself in two quantifiable ways-the time it takes the vehicles to complete their mission and the amount of effort on the part of the operator. Faster and less effort is better. If we use value as the metric then we have capability / cost. In this case, capability is defined as "the

mission to be accomplished" and cost is the time and the effort to accomplish it. Value can increase either by accomplishing more with a given amount of time and / or effort, or by accomplishing the same amount in less time and / or with less effort.

The next step is to develop a quantitative measure of capability, time and effort that are in proportion to their importance. The best approach to this is a normalization process similar to "utility functions" used in decision making: Variables that constitute capability have a lower threshold of zero for any capability that is considered unacceptable (has no worth) and a value of 1 corresponding to the maximum limits of the machine. Variables that constitute cost have a natural lower threshold at zero and have no upper threshold but must be normalized so that one of the costs does not dominate all others. For time we can use the maximum allowable time to complete the mission as a normalization factor and for effort we can use the capacity of the transmission channel, the operator, or the batteries as the basis for normalization - whichever ceiling we hit first.

Realistically, at this point, we can make performance measurements by defining specific mission scenarios and then computing the appropriate metrics. I believe that in general, for any given scenario, greater value will imply greater performance in a statistically significant way. However, this will only be a relative measure unless one can develop some method of absolutely defining the difficulty of the mission, the risk, and some "standard operator" similar to MRT's "standard observer"; even such properties as the value of having physically separate machines in terms of survivability, loss of life, equipment, etc. should be folded into a true operational evaluation. However, at this point, as a technology demonstration, I'm not sure that we know enough to develop defensible metrics nor should we expend excessive resources on trying to develop them, although the obvious place to start is with component performance metrics (bottom's up!).

### 7.2.7 Monica M. Glumm, U.S. Army Human Engineering Laboratory

The development of standardized methods of measurement and performance evaluation are essential to the ARPA UGV community. Performance data derived from relevant and repeatable testing techniques and methodologies are needed to identify and assess viable technologies, and facilitate comparison of the effectiveness of competing UGV systems and their mission modules.

The greater focus of UGV technologies is on machine autonomy, but there remain elements of human / machine interface and interaction that will impact system performance. Assessment of the human's ability to perform his / her supervisory or managerial functions is integral to the assessment of system effectiveness and must be addressed in the development of methods of measuring and evaluating the performance of these technologies. In the event of system failure, there is also a need to assess the operator's ability to intervene and effectively continue the mission. Although technology is advancing by "leaps and bounds" with major emphasis on increased autonomy, there is still a requirement to relate the advantages that these new technologies offer to some existing method or capability.

Some of these new technologies are candidates for incorporation into near and mid-term UGV development efforts. The program managers, who must weigh the costs against the benefits, need to know how much a new technology will buy in terms of enhanced system performance. The soldier-supervisor, who ultimately is responsible for the success or failure of the mission, wants these data expressed in terms that are concrete and relatable to the world in which he / she must operate. If the researcher and the tester are to effectively and efficiently support the ARPA UGV community, they must be responsive to the needs of these customers - the users.

Wherever possible and appropriate, a consistent set of testing techniques, performance measures and evaluation criteria should be developed to assess the performance of both the human and the machine. Commonality in methodologies and measures will:

- enable collection of those data needed to establish a baseline for comparison of performance
- assist in locating and quantifying the strengths and weaknesses within the system
- aid in the allocation of tasks between the man and the machine
- and define the level of autonomy needed to maximize system effectiveness.

These methods and measures must be relatable to the real world. They must also lend themselves to field execution and provide the potential for development of

high-fidelity simulation to assess the performance of tasks that are properly benchmarked and calibrated against field measures. These simulations would examine a wide variety of system parameters and further explore the envelopes of system performance at a reduced cost to the customer and risk to equipment.

# 8. Appendix A - List of Workshop Attendees

| Name | Telephone/Fax |
|---|---|
| Jim Albus<br>NIST<br>Bldg. 220, Rm B124<br>Gaithersburg, MD 20899 | 301/975-3418<br>Fax: 301/990-9688 |
| Dave Anhalt<br>Martin Marietta Defense<br>Space & Communications<br>P.O. Box 179<br>MS H8380<br>Denver, CO 80201 | 303/971-3348<br>Fax: 303/971-4702 |
| Jim Antonisse<br>Mitre Corporation<br>7525 Colshire Dr., #2401<br>McLean, VA 22102 | 703/883-7887 |
| John Baras<br>AIMS, Inc.<br>6159 Executive Blvd.<br>Rockville, MD 20852 | 301/468-5541<br>Fax: 301/468-5543 |
| Bob Bolles<br>SRI<br>SRI International EK290<br>333 Ravenswood Ave.,<br>Menlo Park, CA | 415/859-4620<br>Fax: 415/859-3735 |
| Hal Burke<br>AMSAA<br>AMXSY-CS<br>Aberdeen Proving Grd, MD 21005-5071 | 410/278-4676<br>Fax: 410/278-4694 |
| Peter Burt<br>David Sarnoff Research Ctr.<br>Mail Stop W-357<br>201 Washington Rd.<br>Princeton, NJ 08543-5300 | 609/734-2451<br>Fax: 609/278-4694 |
| Wei Chang<br>Army Research Labs<br>2800 Powder Mill Rd.<br>AMSLC-SS-FA<br>Adelphi, MD 20783 | |
| Daniel DeMenthon | 301/405-1746 |

University of Maryland
Ctr for Automation Research
College Park, MD 20742

Dan Dudgeon
MIT Lincoln Laboratory
244 Wood St., Rm M2005
Lexington, MA 02173
617/981-7216
Fax: 617-981-4094

Edward Durfee
University of Michigan
1101 Beal Ave., ATL Bldg.
Ann Arbor, MI 48109-2110
313/936-1563
Fax: 313/763-1260

Phil Emmerman
Army Research Labs
2800 Powder Mill Rd.
AMSLC-SS-FA
Adelphi, MD 20783
301/394-30
Fax: 301/394-3903

Carl Friedlander
ISX Corporation
4353 Park Terrace Dr.
Westlake Village, CA 91361
818/706-2020
Fax: 818/706-2056

Fred Garrett
Martin Marietta Defense
Space & Communications
PO Box 179, MSH8380
Denver, CO 80201
303/977-4572
Fax: 303/971-4702

Lynne Gilfillan
LGA Inc.
12699 Sabastian Dr.
Fairfax, VA 22030
703/815-2373

Monica Glumm
U.S. Army Human Engineering Lab.
Attn: SLCHE-CS, Bldg. 459
Aberdeen Proving Grd., MD 21005-5001
410/278-5955
Fax: 410/278-8828

Thomas Haduch
U.S. Army Human Eng Lab.
Attn: SLCHE-CS
Aberdeen Proving Grd, MD 21005
410/278-5898
Fax: 410/278-8828

Karan Harbison-Briggs
Univeristy of Texas-Arlington
7300 Jack Newell Blvd. 5
PRISM Lab.
817/794-5900
Fax: 817/794-5952

Ft. Worth, TX 76118

Scott Harmon
Hughes Research Lab.
3011 Malibu Canyon
Malibu, CA 90265
310/317-5140
Fax: 310/317-5695

Steve Hennessy
Martin Marietta
Defense Space & Communications
PO Box 179 MS H8380
Denver, CO 80201
303/977-6541
Fax: 303/971-4702

Martin Herman
NIST
Bldg. 220, Rm B124
Gaithersburg, MD 20899
301/975-3441
Fax: 310/990-9688

Tsai Hong
NIST
Bldg. 220, Rm B124
Gaithersburg, MD 20899
301/975-3441
Fax: 310/990-9688

Charles Jacobus
Cybernet Systems Corp.
1919 Green Rd., Suite B 101
Ann Arbor, MI 48105
313/668-2567
Fax: 313/668-8780

Ramesh Jain
University of Michigan
1101 Beal Ave., ATL Bldg.
Ann Arbor, MI 48109-2110
313/764-8505
Fax: 313/763-1260

Martin Lahart
Night Vision & Electro Optics
Attn: AMSEL-RD-NV-ISPO Directorate
MS 677
Ft. Belvoir, VA 22060-5677
703/704-3471
Fax: 703-704-1705

Jim Leonard
Automatic Target Research
WL/AARA
WPAFB, OH 45433-6543
513/255-1115
Fax: 513/476-4414

Rurik K. Loder
ARL
SLCBR-SE
Aberdeen Proving Grd, MD 21005-5066
410/278-9065
Fax: 410-278-3075

Richard Luhrs
303/977-9585

Martin Marietta
PO Box 179 MS H8380
Denver, CO 80201
Fax: 303/977-9585

Larry Matthies
Jet Propulsion Lab.
MS 107-102
4800 Oak Grove Dr.
Pasadena, CA 91109
818/554-3722
Fax: 818/393-5007

Erik Mettala
ARPA SISTO
3701 N. Fairfax Dr., Rm 767
Arlington, VA 22203
703/696-2219
Fax: 703/696-2202

David Morgenthaler
Martin Marietta Defense
Space & Communications
PO Box 179, MS H8380
Denver, CO 80201
303/977-4200
Fax: 303/971-4702

Scott Myers
Robotic Systems Technology
1110 Business Pkwy South
Westminster, MD 21157
410/876-9200
Fax: 415/876-9470

H. Keith Nishihara
Teleos Research
576 Middlefield Rd.
Palo Alto, CA 94301
415/328-8880
Fax: 415/328-8880

Raymond Resendes
Combat Systems Test Activity
Attn: STECS-AA-AR
Aberdeen Proving Grd., MD 21005-5059
410/278-8645
Fax: 410-278-4308

Edward Riseman
Univ. of Massachusetts
Computer Science Dept.
Rm 213, Lederle GRC Bldg.
Amherst, MA 01003
413/545-2756
Fax: 413/545-1249

Roger Schappell
Martin Marietta Defense
Space & Communications
PO Box 179, XL4370
Denver, CO 80201-0179
303/977-4474
Fax: 303/977-7946

Marcel Schoppers
A.D.S./Booz Allen
415/960-7553
Fax: 415/960-7500

1500 Plymouth
Mt. View, CA 34043

Howard Stern
RVSI
425 Rabro Drive East
Hauppauge, NY 11788
516/273-9700
Fax: 516/273-1167

John Thomas
AMSAA
AMXSX-CS
Aberdeen Proving Grd, MD 21005
410/278-6473
Fax:410/278-4694

Chuck Thorpe
Robotics Institute
Carniege Mellon Univ.
Pittsburgh, PA 15213
412/268-3612
Fax:412/621-1970

Richard Volpe
Jet Propulsion Lab
4800 Oak Grove Dr., MS 198-219
Pasadena, CA 91214
818/354-6328
Fax: 818-393-5007

Russell Watts
Cybernet Systems Corp.
1919 Green Rd., Suite B101
Ann Arbor, MI 48105
313/668-2567
Fax: 313/668-8780

Terry Weymouth
University of Michigan
1101 Beal Ave., ATL Bldg.
Ann Arbor, MI 48109-2110
313/764-3726
Fax: 313/763-1260

Thomas Williams
Amerinex Artificial Intel.
409 Main St.
Amherst, MA 01002
413/256-8941
Fax: 413/253-4203